Aplicação do Método Bootstrap em Diagnósticos de Colinearidade: Resultados Experimentais

Ana Paula Pellegrino Bechelli

Orientador: Prof. Dr. Hermann Gerhard Rohrer

Instituto de Matemática Estatística e Ciência da Computação
UNICAMP

Dezembro-1994

Este exemplar corresponde a redação final da tese devidamente corrigida e defendida pela Sra. Ana Paula Pellegrino Bechelli e aprovada pela Comissão Julgadora.

Campinas, 19 de dezembro de 1994

Prof. Dr. _____
Orientador

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação, UNICAMP, como requisito parcial para obtenção do Título de MESTRE em Estatística

Ao Carlos e aos
meus Pais

Agradecimentos

Sempre achei que esta seria a parte mais fácil, mas agora as palavras somem e sei que não vou conseguir dizer tudo aquilo que eu queria, gostaria de relembrar todos os momentos e as pessoas que de alguma maneira participaram e contribuiram para que este trabalho tivesse um começo, um meio e um fim, cheio de sorrisos, lágrimas, decepções mas também muitas conquistas.

Tudo isso começou quando ainda na ENCE três professores especiais, Djalma, Renato e Zé Carlos mostraram que nem só de números é feita a Estatística. A vontade de aprender mais veio da empolgação que o Renato transmitia em cada conversa. Mas a decisão naquele momento parecia tão difícil, vir morar em Campinas, longe de casa e de tudo aquilo que sempre significou segurança. O incentivo dado pelos meus pais foi grande eles faziam tudo parecer mais fácil e já naquele momento eu podia contar com o Carlos, pelo menos eu não ia ter que passar por tudo sozinha.

A chegada em Campinas foi melhor do que eu esperava, o Jordão e a Lílian foram realmente especiais, fico sempre pensando em como eu poderia retribuir tudo o que eles fizeram durante esse tempo todo de mestrado, mas ainda não descobri uma maneira.

A adaptação longe de casa não foi nada fácil, foram horas reclamando no telefone com a minha mãe, o meu pai sempre ligando, como quem não quer nada, mas mostrando que estava sempre ali pronto para ajudar. As saudades eram grandes, que falta eu senti das minhas irmãs, das conversas, das trocas de roupa e até das brigas. Vocês quatro foram demais!!

A Unicamp era um mundo tão diferente da pequena ENCE por onde eu tinha passado. Mas aos poucos tudo foi se acomodando, o reencontro com o Caxandre e Aninha que são hoje grandes amigos, fui conhecendo novas pessoas Ana, Sílvia, Hildete, Paulo, Damião e tantas outras que tornaram o mestrado bem mais doce. E em cada minuto vivido você, Carlos, estava ali sempre ao meu lado.

O verão passou as aulas regulares começaram e agora mais três pessoas faziam parte desse novo mundo o Hermann, com aquele jeito meio bravo, bem alemão mas que se tornou um amigo, e mais tarde o meu orientador, que contribuiu muito para que essa tese horas bendita, horas maldita chegasse ao fim. A Eliana e o Mauro cada um com o seu jeito ela mais alegre e bricalhona, ele mais tímido e sério mas os dois sempre com um sorriso, hoje são grandes amigos.

Não posso deixar de falar, do pessoal da biblioteca, do cafezinho, da secretaria, em especial

Luci, D.Cida, Cidinha, Marcelo vocês foram sempre nota dez.

Mas, nem só de Unicamp foram feitos esses quase 4 anos, ainda sobrou tempo para as festas, viagens e tudo que acontecia adicionava mais pessoas ao grupo dos que torciam para o essa tão famosa tese chegar ao final, vocês foram todos especias: Nono, Nona, Avós, Tios, Primos, Diva, Vicente, Adri, Léa, Tid e você Rebeca que foi sempre motivo de muita alegria. Enfim, todos aqueles que participaram, um beijo enorme!

E finalmente a você Carlos, meu amigo, amor e companheiro para quem todos os agradecimentos ainda seriam poucos, você esteve sempre ali ao meu lado o tempo todo, mesmo quando tudo parecia que ia dar errado você acreditava e me incentivava. Foram muitos momentos difíceis, mas nós conseguimos. Mas não esqueça que isso foi apenas o começo...

Índice

Capítulo 1

Introdução

Nas últimas décadas o modelo de regressão linear tem sido amplamente utilizado como uma ferramenta da análise quantitativa nas áreas de Ciências Econômicas, Biomédicas e Sociais. A sua vasta utilização está documentada pela farta literatura encontrada acompanhada de programas estatísticos de fácil acesso.

Muitas pesquisas são feitas com o intuito de quantificar a relação entre o comportamento de uma variável resposta $\mathbf{y}$, chamada de variável dependente, e uma matriz de variáveis, supostas linearmente independentes, $\mathbf{X}$. O modelo de regressão linear é definido por:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \text{ com } E(\mathbf{y} \mid \mathbf{X}) = \mathbf{X}\beta$$

onde

$\mathbf{y}$ é o vetor resposta $(n \times 1)$ com suposta distribuição Normal;

$\mathbf{X}$ é a matriz de variáveis explicativas independentes $(n \times p)$;

β é o vetor de parâmetros desconhecidos $(p \times 1)$;

ϵ é o vetor de erros aleatórios independentes $(n \times 1)$ tal que $\epsilon \sim N(0, \mathbf{I}\sigma^2)$, sendo $\mathbf{I}$ a matriz identidade;

e ainda,

n é o número de observações do modelo;

p é o número de parâmetros.

O modelo de regressão linear tem sido tratado na literatura de uma maneira bastante abrangente e eficiente, englobando além da estimativa dos parâmetros, os métodos de verificação da adequabilidade do ajuste do modelo. Estes métodos envolvem análise dos resíduos, detecção de valores extremos e teste F, entre outros. Estas estatísticas refletem quão bem o modelo ajustado estima a resposta observada, mas não necessariamente a validade da predição do modelo para valores da população.

Entretanto, em alguns casos mesmo não tendo sido encontrado nenhum problema nas avaliações mencionadas anteriormente, os resultados obtidos a partir da análise de regressão podem mostrar-se não satisfatórios, com estimativas sem sentido ou irreais, como mostra o exemplo que será apresentado no Capítulo 2, indicando que ainda existe algum problema com os dados que não foi percebido pelo pesquisador. Uma das causas desses problemas pode ser a presença de variáveis colineares no modelo.

Muitas definições de colinearidade aparecem na literatura, algumas mais intuitivas do que objetivas. A definição utilizada nesse trabalho é aquela proposta por Johnston (1962)[13], Silvey (1969)[22] e outros, e consiste em definir a colinearidade a partir do conceito da dependência linear de um conjunto de vetores colunas da matriz $\mathbf{X}$. Uma caracterização de dependência linear pode ser dada da seguinte maneira: os vetores $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ são linearmente dependentes se existem escalares $a_1, a_2, \ldots, a_p$, nem todos iguais a zero, tal que

$$\sum_{j=1}^{p} a_j \mathbf{x}_j = \mathbf{0}. \tag{1.1}$$

Em outras palavras quando a equação (1.1) é valida, a dependência linear na matriz $\mathbf{X}$ existe. Em muitos conjuntos de dados, entretanto, a igualdade (1.1) não vale exatamente, mas sim aproximadamente, isto é, vale uma relação do tipo

$$\inf_{\|\mathbf{a}\|=1} \| \sum_{j=1}^{p} a_j \mathbf{x}_j \| < \eta, \tag{1.2}$$

onde $\|\cdot\|$ é a norma euclidiana, $\eta > 0$ e $\mathbf{a} = (a_1, a_2, \ldots, a_p)$.

Neste caso, através da equação (1.2), mantemos a existência da colinearidade permitindo uma maior flexibilidade ao conceito, isto é, não exige-se mais que a dependência linear seja exata, mas sim que seja limitada por um valor de η arbitrariamente pequeno. Desse modo temos que a questão da colinearidade não reside somente no fato dela existir ou não, mas sim no grau de dependência entre as variáveis.

Durante este trabalho nós utilizamos o termo colinearidade para descrever a relação de dependência aproximada entre as variáveis (expressa conforme a desigualdade 1.2).

Muitas das aplicações da análise de regressão linear tem como principal objetivo as estimativas individuais dos parâmetros. A partir do modelo ajustado, decisões importantes são frequentemente tomadas baseadas na magnitude dessas estimativas individuais ou testes de significância a elas associadas. Além disso, relações de causa e efeito entre a variável resposta e as variáveis explicativas são feitas, entretanto, a presença de variáveis colineares no modelo de regressão pode fazer com que estas inferências sejam confusas ou até mesmo erradas, já que os coeficientes estimados, através do método de Mínimos Quadradros, das variáveis envolvidas em relações de dependência, podem apresentar uma variância grande, gerando muita incerteza sobre os coeficientes estimados.

O problema da colinearidade é frequentemente observado em estudos da área biomédica que envolvem dados não experimentais ou um grande volume de variáveis. Gunst e Mason (1977)[10] utilizaram um estudo sobre ferimentos na espinha dorsal para enfocar as vantagens de examinar a presença de variáveis colineares antes do ajuste do modelo de regressão.

Mason e outros (1975)[18] definiram três possíveis causas da colinearidade:

1) Devido a um modelo super definido;

2) Devido as técnicas de amostragem adotadas;

3) Devido a restrições físicas do modelo ou da população.

Um modelo super definido é aquele em que existem mais variáveis regressoras do que observações. Este tipo de modelo é muito frequente na área biomédica onde muitas informações são requeridas de cada indivíduo em estudo. Para a solução desse problema o estudo deve ser reformulado visando a inclusão de mais indivíduos no estudo; este não é um problema de colinearidade entre as variáveis; ou deve ser feita a retirada das variáveis equivalentes que possam estar gerando o problema da colinearidade. Suponha por exemplo, que a pressão sanguínea de um grupo de pacientes foi medida no braço direito e no braço esquerdo e estas duas variáveis foram utilizadas no modelo. Deixando de lado algum erro de medição que possa ter ocorrido, as observações obtidas a partir dessas variáveis são quase equivalentes e não devem ser utilizadas conjuntamente (não considerando casos patológicos onde por exemplo, uma das artérias esteja obstruída), e a eliminação de uma das variáveis deve ser feita para evitar o problema da colinearidade.

Por outro lado existem situações em que a amostra utilizada não compreende todo o espaço representativo da população, isto é, a amostra retirada representa um subespaço da população de interesse. Neste caso a solução ideal para evitar o problema da colinearidade seria fazer uma nova amostra ou coletar dados adicionais. Entretanto, em muitos casos isto não é possível por razões de custo ou mesmo de viabilidade do projeto. Um exemplo que reflete essa situação pode ocorrer no estudo de uma determinada doença que afeta pessoas de ambos os sexos e de todas as idades, mas na amostragem foram selecionadas apenas mulheres com idade inferior a 30 anos e homens com idade superior a 30 anos: isto é, um subespaço da população de interesse. A análise dos dados pode mostrar a existência de colinearidade entre as variáveis sexo e idade, entretanto, essas variáveis não podem ser consideradas "equivalentes " e a retirada de uma delas do modelo prejudicaria a estimação (ou uma tentativa de extrapolação).

A colinearidade devido a restrições físicas é similar à anterior. A diferença está no fato de que ela existe independente da técnica de amostragem empregada. Este tipo de colinearidade restringe a utilização do modelo e é bastante frequente em estudos

observacionais.

Vimos que em alguns casos não conseguimos solucionar o problema da colinearidade retirando variáveis do modelo, mas é muito importante que saibamos exatamente quais são essas variáveis para não cometermos erros na interpretação individual dos parâmetros, já que nessas situações as estimativas parciais dos coeficientes podem estar bastante distorcidas.

Várias técnicas tem sido utilizadas e outras propostas com o intuito de diagnosticar a colinearidade; a realização desse trabalho tem como objetivo revisar as técnicas clássicas de detecção de colinearidade e introduzir o método de reamostragem bootstrap como uma ferramenta para o diagnóstico.

No Capítulo 2 apresentamos algumas das técnicas mais utilizadas e suas principais deficiências. É dada especial atenção a técnica desenvolvida por Belsley, Kuh e Welsch (BKW; 1980)[1] e detalhada em Belsley (1992)[2], proposta através da fusão dos trabalhos de Kendall-Silvey (1957,1969)[15,22] e de algumas técnicas desenvolvidas por analistas numéricos com o intuito de detectar o mau condicionamento de uma matriz. Estes trabalhos geraram o Número e o Índice de Condição e também a Proporção de Decomposição da Variância (PDV). A habilidade desse conjunto de técnicas em diagnosticar a presença de variáveis colineares e, além disso, detectar quanto a variância de cada variável esta sendo afetada pela presença da colinearidade fez com que muitos trabalhos publicados a partir dessa época utilizassem estes diagnósticos. Como exemplo citamos Montgomery e Peck (1982)[19] e Rawlings (1988)[20].

No Capítulo 3 são apresentados vários modelos de simulação Monte Carlo que foram utilizados para o cálculo do Coeficiente de Variação Estimado, do Número e Índice de Condição, da Proporção de Decomposição da Variância e de Intervalos de Confiança dos estimadores de β, das amostras geradas a partir da utilização do método bootstrap, com o intuito de detectar a presença de variáveis colineares.

No Capítulo 4 apresentamos dois exemplos, o primeiro retirado de Rawlings e o segundo do manual SAS System for Regression (1986)[21]. Empregamos as técnicas de

detecção de colinearidade propostas no Capítulo 3 e, finalmente, analisamos os resultados.

No Apêndice apresentamos os programas utilizados neste trabalho.

Capítulo 2

Diagnósticos Clássicos de Colinearidade

2.1 Introdução

A presença de variáveis colineares no modelo de regressão linear tem sido reconhecida como uma fonte de problemas na estimação, cálculo e interpretação dos parâmetros. Entretanto, a existência de interrelações entre as variáveis explicativas nem sempre é avaliada corretamente.

Já que a análise de um conjunto de dados em relação a existência da colinearidade deve ser feita como um passo inicial em qualquer análise de regressão múltipla, devemos ter disponíveis métodos de diagnosticar colinearidade, bem como métodos para verificar se hipóteses básicas do modelo são violadas.

Antigamente, quando as regressões eram feitas utilizando modelos que envolviam um volume pequeno de dados e parâmetros, era mais fácil detectar visualmente pontos estranhos e algumas formas óbvias de colinearidade a medida que os dados eram analisados. Com a introdução dos computadores e utilização cada vez mais frequente de modelos que envolvem um grande volume de dados e parâmetros, a análise (visual) desses dados por parte do pesquisador à procura da existência de colinearidade fica cada vez mais difícil.

A não utilização de diagnósticos que possibilitem a detecção de variáveis colineares e consequentemente a aplicação de alguma medida corretiva ou pelo menos o entendimento

da estrutura apresentada pelos dados, pode fazer com que a sensibilidade das estimativas dos parâmetros do modelo fique prejudicada.

Diagnósticos de colinearidade eficientes e objetivos são necessários para um bom entendimento da estrutura do conjunto de dados que será analisado. Neste capítulo apresentamos as técnicas clássicas mais utilizadas para a detecção da colinearidade

2.2 Diagnósticos de Colinearidade

Muitos procedimentos tem sido utilizados com o intuito de diagnosticar a colinearidade, porém algumas falhas puderam ser observadas. Mostramos a seguir alguns desses diagnósticos.

Gunst (1983)[12] utiliza um exemplo retirado de Gunst e Mason (1980)[11] para mostrar o problema da colinearidade. Utilizamos este exemplo para mostrar que alguns fenômenos são frequentemente apontados como decorrentes da colinearidade, embora nem sempre isto seja verdade. Os dados analisados por Gunst (1983)[12] pretendem construir uma equação de predição para uma relação linear existente entre altura (HEIGT) e o seguinte conjunto de variáveis explicativas: altura sentado (SITHT), comprimento da parte superior do braço (UARM), comprimento do antebraço (FORE), comprimento da mão (HAND), comprimento da parte superior da perna (ULEG), comprimento da parte inferior da perna (LLEG), comprimento do pé (FOOT), comprimento Braquial (BRACH) e o comprimento Tibio-Femural (TIBIO).

A tabela 2.1 mostra a análise de variância para a estimação dos coeficientes feita por Mínimos Quadrados.

Tabela 2.1

		Análise de Variância			
Fonte de Variação	**Graus de Liberdade**	**Soma de Quadrados**	**Média de Quadrados**	F -	R^2 -
Regressão	9	683.824	75.980	21.272	0.893
Erro	23	82.152	3.572	-	-
Total	32	765.976	-	-	-

O valor da estatística F e o coeficiente de determinação R^2, que mostra quanto da variabilidade de **y** está sendo modelada pelas variáveis regressoras, indicam que o modelo pode ser considerado adequadamente ajustado aos dados. Os critérios tipicamente utilizados para medir a qualidade do ajuste de modelos de regressão indicam que o ajuste por Mínimos Quadrados para este conjunto de dados parece satisfatório. Entretanto algumas dificuldades aparecem quando a interpretação dos coeficientes estimados é feita.

Uma característica estranha dos coeficientes estimados mostrada na tabela 2.2 é a ocorrência de valores negativos. Espera-se que todos os coeficientes utilizados para explicar **altura** sejam não negativos, pois todas as variáveis explicativas medem crescimento. Além disso, apesar do ajuste do modelo ter sido considerado bom, o erro-padrão da maioria dos coeficientes estimados é grande, resultando em valores não significantes para as estatísticas t (de Student).

Tabela 2.2

Variável Preditora	Estimativa de M.Q.	Erro Padrão	Estatística t
SITHT	0.798	0.156	5.105
UARM	0.459	4.156	0.110
FORE	-0.454	4.771	-0.095
HAND	0.795	0.549	1.448
ULEG	-0.758	4.248	-0.178
LLEG	2.147	4.517	0.475
FOOT	0.936	0.669	1.338
BRACH	0.264	1.573	0.168
TIBIO	-0.445	1.872	-0.238

A ocorrência desses valores é constantemente apontada como uma consequência da colineridade, apesar de não ser necessariamente a presença da colinearidade a causa destas distorções. Amostras pequenas, dados não representativos, um modelo mau especificado, entre outros fatores, podem contribuir para a ocorrência dessas distorções.

Uma outra medida bastante utilizada para avaliar a presença da colinearidade é a matriz de correlação, $\mathbf{R}$, entre as variáveis $\mathbf{X}$, entretanto existem também algumas deficiências. Apesar de um coeficiente de correlação alto entre duas variáveis ser um indicativo da presença da colinearidade, o contrário não implica que a colinearidade não esteja presente. É possível que três ou mais variáveis sejam colineares mesmo que os pares não apresentem uma alta correlação. Mansfield e Helms (1982)[16] exibiram um exemplo que mostra que pode exister colinearidade entre as variáveis explicativas mesmo não existindo correlação alta entre os pares de variáveis.

Assumindo que a matriz $\mathbf{X}$ esteja centralizada e normalizada para cada vetor-coluna ter tamanho 1, então $\mathbf{R}^{-1} = (\mathbf{X}^t\mathbf{X})^{-1}$. Os elementos da diagonal de $\mathbf{R}^{-1}$, são chamados

de Fator de Inflação da Variância, FIV, Chatterjee e Price (1977)[3], e expresso por:

$$FIV_i = \frac{1}{1 - R_i^2} \tag{2.1}$$

onde R_i^2 é o coeficiente de correlação múltipla de $\mathbf{x}_i$ regredido com as demais variáveis explicativas. O termo Fator de Inflação da Variância deriva do fato de que a variância do i-ésimo coeficiente de regressão, $Var(\hat{\beta}_i)$, obedece a seguinte relação

$$Var(\hat{\beta}_i) = \frac{Var(\epsilon)}{\mathbf{x}_i^t \mathbf{x}_i} FIV_i$$

Ou seja, um alto valor de FIV_i indica, pela equação (2.1), que R_i^2 esta perto de um, apontando desse modo a presença de variáveis colineares. Além disso, quanto maior o valor do FIV_i, mais inflacionada estará a $Var(\hat{\beta}_j)$; sendo que, quando as variáveis são ortogonais o valor de FIV_i é igual a 1 para todo i.

Apesar deste ser um diagnóstico computacionalmente fácil e frequentemente encontrado na literatura, ver por exemplo, Montgomery e Peck (1982)[19] e Rawlings (1988)[20], sua falha consiste em não diagnosticar colinearidades que envolvam mais de duas variáveis, pois foi visto que este diagnóstico está baseado na inversa da matriz de correlação $\mathbf{R}$, que como foi dito anteriormente, nem sempre é capaz de identificar a presença da colinearidade.

Farrar e Glauber (1967)[7] desenvolveram um indicador da existência de colinearidade baseado na hipótese de que a matriz de dados $\mathbf{X}_{n \times p}$ é uma amostra de tamanho n de uma distribuição *Normal p − variada.* Sob a hipótese de que as colunas de $\mathbf{X}$ são ortogonais, afirmaram que uma transformação do determinante da matriz de correlação $\mathbf{R}$ é aproximadamente distribuída como uma χ^2, produzindo desse modo uma medida do afastamento dos dados em relação a ortogonalidade. Haitovsky (1969)[14] criticou a medida por estar baseada no afastamento da ortogonalidade, o que frequentemente indicava a presença da colinearidade quando nenhum problema prático na verdade existia.

Kendall (1957)[15] e Silvey (1969)[22] sugeriram a utilizacão dos autovalores e au-

tovetores da matriz $\mathbf{X}^t\mathbf{X}$ como um diagnóstico de colinearidade, partindo do princípio de que a existência de uma dependência linear entre os dados da matriz $\mathbf{X}$ é refletida em pelo menos um autovalor nulo ou igual a zero. Assumem que valores pequenos para os autovalores é um indicativo de colinearidade. Neste caso o problema consiste em avaliar o que seria um autovalor pequeno.

Nenhuma dessas técnicas foi completamente bem sucedida em diagnosticar a presença de variáveis colineares.

Os analistas numéricos, interessados nas propriedades de condicionamento de uma matriz $\mathbf{A}$ (de um sistema de equacões lineares $\mathbf{Az} = \mathbf{c}$) que permita que as solucões para $\mathbf{z}$ sejam obtidas com certa estabilidade numérica, desenvolveram um conjunto de ferramentas muito úteis para uma análise da presença da colinearidade. A relação existente entre a solução de um sistema de equações lineares e o problema da colinearidade deve-se ao fato de que o estimador de mínimos quadrados é a solução do sistema linear

$$(\mathbf{X}^t\mathbf{X})\beta = \mathbf{X}^t\mathbf{y} \tag{2.2}$$

e a presença da colinearidade faz com que a matriz $\mathbf{X}^t\mathbf{X}$ seja mau condicionada.

Os diagnósticos propostos por BKW uniram as técnicas desenvolvidas pelos analistas numéricos e os trabalhos de Kendall e Silvey de maneira a produzir um conjunto de índices capazes de diagnosticar todas as variáveis envolvidas em colinearidade e como a variância estimada de cada coeficiente está sendo afetada pela colinearidade. Embora BKW afirmem que estes diagnósticos não envolvem novos conceitos, a novidade consiste na sua combinação feita de maneira a ajudar o usuário de regressão linear a resolver dois problemas de diagnóstico: a detecção e o acesso ao dano causado pela colinearidade.

Na secção seguinte é apresentada a decomposição em valores singulares associada à noção de condicionamento da matriz $\mathbf{X}$ e a decomposição da variância estimada em relação aos valores singulares. A combinação desses conceitos gerou os diagnósticos propostos por BKW:

1) Número e o índice de condição.

2) A proporção de decomposição da variância.

2.2.1 Número de Condição e Índice de Condição

A noção de número de condição veio através dos trabalhos desenvolvidos pelos analistas numéricos que estavam interessados em determinar a solução de um sistema de equações lineares do tipo $\mathbf{Az} = \mathbf{c}$,onde $\mathbf{A}_{n\times n}$ é uma matriz não singular, e observar qual seria o efeito no vetor de solucões $\mathbf{z}$ para pequenas mudancas em $\mathbf{A}$ ou $\mathbf{c}$. O número de condição de uma matriz produz um indicador da sensibilidade da solucão $\mathbf{z} = \mathbf{A}^{-1}\mathbf{c}$ em respeito às distorções em $\mathbf{A}$ ou $\mathbf{c}$. Grandes variações encontradas no vetor de soluções $\mathbf{z}$ com pequenas mudanças em $\mathbf{A}$ ou $\mathbf{c}$ são um indicativo de que a matriz $\mathbf{A}$ é mau condicionada.

Uma das fontes de mau condicionamento de uma matriz está asssociada à presença da colinearidade entre as variáveis explicativas, fazendo com que a estimação através de mínimos quadrados, isto é, $\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$, fique muito instável.

Para entender a relevância do número de condição como diagnóstico de colinearidade vemos como é por ele detectado o mau condicionamento de uma matriz. Como o número de condicão depende da norma utilizada, utilizamos a norma espectral que é uma generalização da norma euclidiana.

Definição: Seja $\mathbf{A}$ uma matriz $n \times n$. A norma espectral de $\mathbf{A}$, denotada por $\| \mathbf{A} \|$, é definida como:

$$\| \mathbf{A} \| \equiv \sup_{\|\mathbf{z}\|=1} \| \mathbf{Az} \| \tag{2.3}$$

sendo a norma utilizada do lado direito a norma euclidiana.

Pode ser mostrado também que $\| \mathbf{A} \| = \mu_1$, isto é, o seu maior valor singular e a $\| \mathbf{A}^{-1} \| = \frac{1}{\mu_p}$, sendo μ_p o seu menor valor singular (a serem definidos a seguir).

Seguindo o procedimento de Forsythe e Moler (1967)[8], supomos que os valores de $\mathbf{A}$ e $\mathbf{c}$ estão sujeitos a uma incerteza, e estamos interessados em saber qual o efeito dessa incerteza na solucão do vetor $\mathbf{c}$. Supomos, primeiramente que $\mathbf{A}$ é conhecida exatamente

e que $\mathbf{c}$ está sujeito a uma incerteza δ, por exemplo $\mathbf{c} + \delta\mathbf{c}$ gerando $\mathbf{A}(\mathbf{z} + \delta\mathbf{z}) = \mathbf{c} + \delta\mathbf{c}$. Vemos, então, quão grande pode ser $\delta\mathbf{z} = \mathbf{A}^{-1}\delta\mathbf{c}$, isto é, qual será a mudança em $\mathbf{z}$ gerada pela mudança feita em $\mathbf{c}$.

Através das propriedades de uma norma de matriz ver, p.e., Forsyth e Moler (1967)[8] temos que: (as normas dos vetores são as normas euclidianas)

$$\| \delta\mathbf{z} \| \leq \| \mathbf{A}^{-1} \| \cdot \| \delta\mathbf{c} \| \tag{2.4}$$

Temos também que $\mathbf{c} = \mathbf{A}\mathbf{z}$, implicando em

$$\| \mathbf{c} \| \leq \| \mathbf{A} \| \cdot \| \mathbf{z} \| \tag{2.5}$$

Multiplicando (2.4) por (2.5) temos:

$$\| \delta\mathbf{z} \| \cdot \| \mathbf{c} \| \leq \| \mathbf{A} \| \cdot \| \mathbf{A}^{-1} \| \cdot \| \mathbf{z} \| \cdot \| \delta\mathbf{c} \| \tag{2.6}$$

De onde segue que:

$$\frac{\| \delta\mathbf{z} \|}{\| \mathbf{z} \|} \leq \| \mathbf{A} \| \cdot \| \mathbf{A}^{-1} \| \frac{\| \delta\mathbf{c} \|}{\| \mathbf{c} \|} \tag{2.7}$$

A partir da equação (2.7) define-se o número de condicão, $K(\mathbf{A})$, para qualquer matriz não singular $\mathbf{A}$ como $K(\mathbf{A}) = \| \mathbf{A} \| \cdot \| \mathbf{A}^{-1} \|$, ou seja

$$\frac{\| \delta\mathbf{z} \|}{\| \mathbf{z} \|} \leq K(\mathbf{A}) \frac{\| \delta\mathbf{c} \|}{\| \mathbf{c} \|} \tag{2.8}$$

Agora $\frac{\|\delta\mathbf{c}\|}{\|\mathbf{c}\|}$ pode ser interpretado como uma medida da incerteza relativa do vetor $\mathbf{c}$ e $\frac{\|\delta\mathbf{z}\|}{\|\mathbf{z}\|}$ como uma medida da incerteza relativa do vetor $\mathbf{z}$, devido a incerteza em $\mathbf{c}$. Desenvolvimento análogo pode ser feito para o caso em que a matriz $\mathbf{A}$ está sujeita a incerteza.

Portanto o número de condição de $\mathbf{A}$ será definido como:

$$K(\mathbf{A}) = \frac{\mu_1}{\mu_p} \geq 1$$

ou

$$K(\mathbf{X}^t\mathbf{X}) = \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{1/2} \geq 1 \tag{2.9}$$

onde λ é o autovalor da matriz $\mathbf{X}^t\mathbf{X}$.

O número de condicão para uma matriz $\mathbf{X}$ ortonormal, ou seja $\mathbf{X}^t\mathbf{X} = \mathbf{I}$, é igual a 1. Por outro lado a singularidade de $\mathbf{X}^t\mathbf{X}$ implica em $\lambda_{\min} = 0$, então, $K(\mathbf{X}^t\mathbf{X}) \to \infty$. Desse modo temos que o número de condição varia no intervalo de $[1, \infty)$.

Em conecção com a definição de número de condição, podemos definir :

$$\eta_k = \frac{\mu_{\max}}{\mu_k}, \ k = 1, \ldots, p \tag{2.10}$$

como o $k - \acute{e}simo$ índice de condiçãode uma matriz $\mathbf{X}_{n\times p}$. Sendo $\eta_k \geq 1$ para todo k. O maior valor de η_k é também o número de condição da matriz $\mathbf{X}$. Através dos índices de condição pode-se determinar quantas relações de dependência existem entre as colunas da matriz $\mathbf{X}$, isto é, cada valor alto para o índice de condição resulta em uma dependência entre as colunas da matriz $\mathbf{X}$.

O cálculo do número e índices de condição e a proporção de decomposição da variância, que será definida na próxima secção, estão diretamente ligado a decomposição em valores singulares (DVS).

2.2.2 Decomposição em Valores Singulares

Os diagnósticos propostos por BKW utilizam a DVS da matriz $\mathbf{X}_{n\times p}$ que está relacionada com o auto-sistema de $\mathbf{X}^t\mathbf{X}$ e com a análise de dependências lineares.

Qualquer matriz $\mathbf{X}_{n\times p}$ (n observações e p variáveis explicativas) pode ser decomposta em, ver p.ex. Golub (1969)[9],

$$\mathbf{X} = \mathbf{UDV}^t, \tag{2.11}$$

onde $\mathbf{U}^t\mathbf{U} = \mathbf{V}^t\mathbf{V} = \mathbf{I}$ e $\mathbf{D}$ é uma matriz diagonal não negativa com elementos μ_k, $k = 1, 2, ..., p$, chamados de valores singulares de $\mathbf{X}$. A relação (2.11) vale independente da matriz $\mathbf{X}$ estar normalizada ou centralizada, mas para efeito dos diagnósticos de colinearidade é necessário que as colunas da matriz $\mathbf{X}$ estejam normalizadas.

A relacão entre a DVS da matriz $\mathbf{X}$ e o auto-sistema da matriz $\mathbf{X}^t\mathbf{X}$ vem do seguinte fato:

$$\begin{aligned} \mathbf{X}^t\mathbf{X} &= (\mathbf{UDV}^t)^t(\mathbf{UDV}^t) \\ &= (\mathbf{VDU}^t)(\mathbf{UDV}^t) \\ &= \mathbf{VD}^2\mathbf{V}^t \end{aligned}$$

Já que $\mathbf{V}$ é uma matriz ortonormal que diagonaliza $\mathbf{X}^t\mathbf{X}$, segue que os elementos da diagonal de $\mathbf{D}^2$ são os autovalores reais da matriz simétrica $\mathbf{X}^t\mathbf{X}$ e as colunas ortonormais de $\mathbf{V}$ são os autovetores de $\mathbf{X}^t\mathbf{X}$.

Consequentemente, a DVS produz informacões que incluem as informacões dadas pelo auto-sistema de $\mathbf{X}^t\mathbf{X}$, entretanto BKW utilizam em seus diagnósticos a decomposicão em valores singulares e afirmam, que do ponto de vista prático, a DVS é mais interessante pois é aplicada diretamente na matriz $\mathbf{X}$. Além disso, argumentam que a noção de número de condicão que é utilizada como diagnóstico da colinearidade foi definida em termos dos valores singulares da matriz $\mathbf{X}$ e não em termos da raiz quadrada dos autovalores de $\mathbf{X}^t\mathbf{X}$. Apontam ainda que apesar de o auto-sistema de $\mathbf{X}^t\mathbf{X}$ e a decomposição em valores singulares de $\mathbf{X}$ serem matematicamente equivalentes, computacionalmente eles não os são, pois os algoritmos desenvolvidos para o cálculo da decomposição em valores singulares permitem que a decomposicão de $\mathbf{X}$ seja feita com uma estabilidade numérica maior à que é possível ao se calcular o auto-sistema de $\mathbf{X}^t\mathbf{X}$, particularmente no caso de interesse, em que temos a matriz $\mathbf{X}$ mau condicionada.

Mostraremos agora qual a relevância da DVS para a determinacão das variáveis en-

volvidas em colinearidade. Assumimos que existe dependência linear entre as colunas de $\mathbf{X}$. Isto implica em *posto* $(\mathbf{X}) = r < p$. Na DVS de $\mathbf{X}$ temos que $\mathbf{U}$ e $\mathbf{V}$ são ortonormais e consequentemente devemos ter *posto*($\mathbf{X}$) =*posto*($\mathbf{D}$). Então existem tantos zeros ao longo da diagonal de $\mathbf{D}$ quanto for a nulidade de $\mathbf{X}$ (determinada pelo número de valores singulares igual a zero), isto é,

$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_r \geq \mu_{r+1} = \mu_{r+2} = \cdots \mu_p = 0$$

Podemos particionar a DVS da seguinte forma:

$$\mathbf{X} = \mathbf{UDV}^t = \mathbf{U} \begin{bmatrix} \mathbf{D}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}^t \tag{2.12}$$

onde $\mathbf{D}_{11}$é uma matriz não singular de ordem $r \times r$. Multiplicando a equação (2.12) por $\mathbf{V}$ temos:

$$\mathbf{XV} = \mathbf{U} \begin{bmatrix} \mathbf{D}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}^t\mathbf{V}$$

Particionando novamente temos:

$$\mathbf{X}[\mathbf{V}_1\mathbf{V}_2] = [\mathbf{U}_1\mathbf{U}_2] \begin{bmatrix} \mathbf{D}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{2.13}$$

de tal modo que $\mathbf{V}_1(p \times r)$,$\mathbf{V}_2(p \times (p-r))$,$\mathbf{U}_1(n \times r)$,$\mathbf{U}_2(n \times (p-r))$,
resultando em duas equacões:

$$\mathbf{XV}_1 = \mathbf{U}_1\mathbf{D}_{11} \tag{2.14}$$

e

$$\mathbf{XV}_2 = \mathbf{0} \tag{2.15}$$

A equacão (2.15) mostra todas as dependências lineares de $\mathbf{X}$.

Então, se $\mathbf{X}$ possui $p - r$ relações lineares entre sua colunas, existem também $p - r$ valores singulares zeros em $\mathbf{D}$ e as variáveis envolvidas em cada uma dessas dependências será determinada pelos elementos não zero de $\mathbf{V}_2$.

No conceito de colinearidade (ver desigualdade (1.2)), as interrelacões entre as colunas de $\mathbf{X}$ não são dependências lineares exatas. Desse modo elementos zeros para os valores singulares ou para os elementos de $\mathbf{V}_2$ são difíceis de ocorrer. Assim, fica difícil determinar a quase nulidade de $\mathbf{X}$ (através dos $\mu's \approx 0$) ou as colunas de $\mathbf{X}$ que não estão envolvidas em relacões de colinearidade (determinada pelos valores não nulos de $\mathbf{V}_2$). Temos, então, que cada quase-dependência entre as colunas de $\mathbf{X}$ se manifesta através de um valor singular pequeno (ficando em aberto o que é "pequeno"). Este ponto de vista corresponde a noção de Silvey (1969)[22], de que a presença da colinearidade seria revelada pela presença de autovalores pequenos. A questão agora é determinar o que pode ser considerado um autovalor pequeno. É através do número de condicão da matriz $\mathbf{X}$ que BKW determinaram a magnitude desse valor.

Para a aplicacão do número e do índice de condição como diagnósticos de colinearidade, nós precisamos que os dados estejam normalizados, i.é., cada vetor da matriz $\mathbf{X}$ deve ter tamanho um, mas não precisam estar centralizados. A normalização em conjunto com a centralização das variáveis para utilização dos diagnósticos propostos por BKW vem sendo motivo de discordância na literatura: Marquardt e Snee (1975)[17] recomendam a centralizacão para evitar o problema de variáveis com origens diferentes, e a normalização é recomendada pois evita problemas de instabilidade numérica em relacão a inversa de $\mathbf{X}^t\mathbf{X}$, quando as variáveis explicativas tem escalas ou magnitudes muito diferentes.

BKW recomendam a normalização das variáveis explicativas para melhorar a estabilidade numérica da matriz $\mathbf{X}$. Argumentam que matrizes que diferem entre si somente pela escala designada para as suas colunas, produzem diferentes índices de condicão. Uma simples mudança na unidade de medida das variáveis, por exemplo de polegadas

para milímetros, que não representa uma mudança na estrutura do modelo, afeta as propriedades numéricas da matriz e resulta em diferentes DVS e consequentemente em diferentes índices de condicão. Além disso, justificam que se os dados são relevantes para o modelo com o termo constante, a matriz $\mathbf{X}$ deve ser não centralizada pois a utilizacão de dados centralizados pode mascarar a funcão da constante em alguma relacão de dependência e produzir resultados de diagnósticos confusos.

Segundo Gunst (1983)[12] a centralizacão das variáveis independentes é recomendada a não ser que o interesse seja diretamente relacionado à colinearidade envolvendo o termo constante.

A utilizacão em conjunto da normalização e centralizacão das variáveis ainda é um assunto em discussão. Optamos por seguir a sugestão de BKW, já que os diagnósticos que apresentamos foram por eles desenvolvidos.

2.2.3 Proporção de Decomposição da Variância

A partir do trabalho desenvolvido por Silvey (1969)[22], BKW mostram que a variância estimada de cada coeficiente da regressão pode ser decomposta, através da DVS, em uma soma de termos onde cada termo está associado a somente um valor singular, produzindo uma maneira de determinar a extensão com que cada dependência está afetando cada variância.

Temos que a matriz de variância-covariância do estimador de Mínimos Quadrados $\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$, é dada por:

$$Var(\hat{\beta}) = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$$

onde σ^2 é a variância comum dos ϵ em um modelo de regressão linear $\mathbf{Y} = \mathbf{X}^t\beta+\epsilon$ (ver, p.ex., Draper & Smith, (1981) [4]). Para o cálculo da PDV vamos considerar a matriz $\mathbf{X}$ normalizada.

Utilizando a DVS, a matriz de variância-covariância de $\hat{\beta}$ pode ser escrita como:

$$\begin{aligned} Var(\hat{\beta}) &= \sigma^2(\mathbf{X}^t\mathbf{X})^{-1} \\ &= \sigma^2[(\mathbf{UDV}^t)^t(\mathbf{UDV}^t)]^{-1} \\ &= \sigma^2[\mathbf{VD}^{-2}\mathbf{V}^t] \end{aligned}$$

ou seja, para o k-ésimo componente de $\hat{\beta}$,

$$Var(\hat{\beta}_k) = \sigma^2 \sum_{j=1}^{p} \frac{\nu_{kj}^2}{\mu_j^2} \tag{2.16}$$

onde μ_j são os valores singulares e ν_{jk} são os componentes dos autovetores. Assim obtemos a decomposição prometida da $Var(\hat{\beta}_k)$, i.é., uma soma onde cada componente do autovetor está associado a somente um dos p valores singulares.

A proporção da decomposição da variância pode ser calculada através das seguintes proporções.

$$\phi_{kj} = \frac{\nu_{kj}^2}{\mu_j^2}$$

e

$$\phi_k = \sum_{j=1}^{p} \phi_{kj}, k = 1, 2, ..., p$$

Então a proporção de decomposição da variância é definida como:

$$\pi_{jk} = \frac{\phi_{kj}}{\phi_k}, k, j = 1, 2, ..., p \tag{2.17}$$

A Tabela 2.3 mostrada a seguir facilita muito a investigação sobre a presença de variáveis colineares através da utilização da proporção de decomposição da variância.

Tabela 2.3

Valores	Proporcão da			
Singulares	$Var(\hat{\beta}_1)$	$Var(\hat{\beta}_2)$	$\cdots$	$Var(\hat{\beta}_p)$
μ_1	π_{11}	π_{12}	$\cdots$	π_{1_p}
μ_2	π_{21}	π_{22}	$\cdots$	π_{2_p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
μ_p	π_{p1}	π_{p2}	$\cdots$	π_{pp}

Como duas ou mais variáveis são necessárias para ser criada uma dependência, ao analisar a tabela 2.3 à procura da colinearidade o pesquisador deve buscar pelo menos dois valores altos de π_{ij}, para $i \neq j$ associados ao mesmo valor singular.

A partir destes diagnósticos, BKW sugerem que uma maneira apropriada de diagnosticar a colinearidade seria observar essas duas condições:

(1) Um valor singular que apresenta um alto índice de condição associado a,

(2) altas proporções de decomposição da variância para duas ou mais variâncias dos coeficientes de regressão estimados.

O número de índices de condição (> 30) no item (1) identifica o número de colinearidades entre as colunas da matriz **X**. Além disso, a determinação, no item (2), de proporções de decomposição da variância (> 0.5) associadas a altos índices de condição identificam as variáveis envolvidas nas correspondentes quase-dependências, ou colinearidade.

Segundo BKW, pode ser determinado através de procedimentos empíricos que um número de condição menor que 15 seria um indicativo de colinearidade fraca entre as variáveis, representando uma correlação menor que 0.9. Um número de condição entre 15 e 30 está associado a uma correlação de 0.9 e números de condição maiores que 30 estão associados a correlações maiores que 0.9 que seria um indicativo de colinearidade forte, considerando que colinearidade forte é aquela que prejudica a estimação através

dos Mínimos Quadrados, levando à distorções nas estimativas e prejudicando possíveis tentativas de extrapolações. Entretanto eles afirmam que esses limites são usuais em modelos econométricos e podem variar dependendo do estudo em questão.

Capítulo 3

Diagnósticos de Colinearidade Através da Utilização do Método Bootstrap

3.1 Introdução

O método bootstrap foi introduzido por Efron (1979)[5], como um método computacional para a estimação do erro padrão da estimativa de um parâmetro, $\hat{\theta}$, isto é, sustentado pela idéia de que, através do bootstrap, a estimação do erro padrão poderia ser feita sem a necessidade de nenhum cálculo teórico, independente de quão matematicamente complicado pudesse ser o estimador $\hat{\theta} = T(\mathbf{X})$, onde $\mathbf{X}$ é um vetor ou uma matriz de observações. É um método computacionalmente intensivo e por isso o seu desenvolvimento e aplicação tem sido ampliados pelo avanço dos recursos computacionais. O método pode ser considerado como uma ferramenta para soluções numéricas de diversos problemas estatísticos que muitas vezes são analiticamente inviáveis, como por exemplo para o cálculo de correlações, intervalos de confiança, entre outros.

Em outras palavras, o bootstrap é um método de reamostragem que tem como objetivo indicar as propriedades amostrais, porém, restrito a uma única amostra, de uma determinada estatística. Denotamos por

$$P \rightarrow \mathbf{x} \tag{3.1}$$

o fato de uma função de probabilidade acumulada (f.p.a.) P, desconhecida, produzir uma amostra $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ através de um mecanismo aleatório. A partir dessa amostra $\mathbf{x}$ calculamos a estatística de interesse $T(\mathbf{x})$, como por exemplo, os estimadores de Mínimos Quadrados dos coeficientes de regressão ou seus intervalos de confiança.

A versão bootstrap para o mecanismo (3.1) é análoga e direta. Tendo a amostra $\mathbf{x}$, gerada pelo mecanismo aleatório relativo à P, obtemos a distribuição empírica $\hat{P}$, como uma estimativa da f.p.a. P. Por amostragem aleatória em $\hat{P}$, geramos uma amostra $\mathbf{x} = (x_1^*, x_2^*, \ldots, x_n^*)$, chamada amostra bootstrap, e denotamos esse fato por

$$\hat{P} \rightarrow x^* \tag{3.2}$$

e do mesmo modo calculamos $T(\mathbf{x}^*)$, a estatística de interesse.

Conceitualmente, o ponto crucial do método bootstrap é reconhecer a analogia entre $P \rightarrow \mathbf{x}$ e $\hat{P} \rightarrow \mathbf{x}^*$. E sua grande vantagem é a possibilidade de gerar as replicações $\mathbf{x}^*$ um grande número de vezes, para estimar as características amostrais de $T(\mathbf{x})$.

Uma revisão das técnicas bootstrap e suas aplicações foi recentemente apresentada por Efron e Tibshirani (1993)[6].

Neste capítulo apresentamos a utilização do método bootstrap como uma ferramenta para o diagnóstico de colinearidade. São propostos quatro diagnósticos de colinearidade e estudamos sua sensibilidade em diagnosticar a presença de variáveis colineares. Estudamos primeiro um modelo de regressão linear com cinco variáveis e 50 observações, onde duas delas são colineares em vários graus e a seguir um modelo com três variáveis colineares.

3.2 Bootstrap em Regressão

No modelo de regressão usual, $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ é o vetor $n \times 1$ da variável resposta, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n)$ é a matriz $n \times p$ de variáveis explicativas. Temos que o modelo de regressão linear é

$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$

onde ϵ o vetor $n \times 1$ de erros não correlacionados e normalmente distribuido tendo média zero e variância $\sigma^2 I$ e β é o vetor de parâmetros desconhecidos $p \times 1$, para os quais o estimador de Mínimos Quadrados é dado por

$$\hat{\beta} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{y}.$$

Existem duas maneiras de utilizar o bootstrap na regressão e a sua escolha depende exclusivamente das características das variáveis explicativas que são impostas. Se as variáveis explicativas são tratadas como fixas, como em um desenho de experimento, a reamostragem é feita no vetor de resíduos que resta como a única parte aleatória do modelo, ver p.ex. Srivastava e Sen (1990)[23]. Entretanto, se as variáveis explicativas são tratadas como aleatórias bem como o vetor resposta $\mathbf{y}$, a reamostragem deve ser feita no par $(\mathbf{y}, \mathbf{X})$ com o intuito de preservar a característica aleatória do modelo. No nosso trabalho consideramos $\mathbf{X}$ como uma variável aleatória e consequentemente utilizamos a reamostragem no par $(\mathbf{y}, \mathbf{X})$.

A aplicação do bootstrap em modelos de regressão onde as variáveis explicativas são tratadas como aleatórias pode ser definido como:

Seja $\mathbf{z}_i = (y_i, x_{i1}, \ldots, x_{ip})$ $i = 1, \ldots, n$ o vetor $1 \times (p+1)$ dos valores associados a i-ésima observação retiramos então, uma amostra de $(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)$ com reposição, onde cada par tem probabilidade $\frac{1}{n}$ de ser selecionado, e denotamos por $(\mathbf{z}_1^*, \mathbf{z}_2^*, \ldots \mathbf{z}_n^*)$ a amostra bootstrap onde $z_i^* = (y_i^*, x_{i1}^*, \ldots, x_{ip}^*)$. Assim é definido uma nova amostra bootstrap $\mathbf{y}^*$ e $\mathbf{X}^*$.

Calculamos então, para cada nova amostra bootstrap, o estimador de mínimos quadrados para dos coeficientes

$$\hat{\beta}^* = \left(\mathbf{X}^{*t}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*t}\mathbf{y}^*$$

O procedimento deve ser repetido quantas forem as replicações desejadas, no nosso trabalho utilizamos as seguintes replicações $B_r \in (250, 397, 630, 1000, 1587, 2520 \text{ e } 4000)$, sendo $B_{r+1} \simeq 1.59 \times B_r$, $r = 1, \ldots, 6$ estes tamnhos de replicações foram escolhidos de tal forma que a visão logaritima nos gráficos fosse facilitada e que também houvesse uma abrangência dos tamanhos de replicações.

3.3 Diagnósticos de Colinearidade

No Capítulo 2 fizemos uma revisão dos diagnósticos de colinearidade propostos e algumas de suas deficiências. Apontamos também os que hoje são mais utilizados, como o número e o índice de condição e a proporção de decomposição da variância.

Vamos apresentar agora quatro diagnósticos, um baseado no coeficiente de variação para cada amostra boostrap, outros dois propostos por BKW só que calculados para cada reamostragem e por último um diagnóstico baseado na razão entre a amplitudes dos intervalos de confiança calculados pelo método percentil, ver Efron e Tibshirani (1993)[6].

Inicialmente, utilizamos os quatro diagnósticos de colinearidade que são apresentados a seguir e verificamos qual a sensibilidade de cada um em diagnosticar a presença de variáveis colineares em um modelo de regressão linear com duas variáveis colineares.

3.3.1 Coeficientes de Variação

A partir dos coeficientes estimados bootstrap, denotados por $\hat{\beta}_k^{*b}$ $k = 0, \ldots, p$, calculamos o coeficiente de variação estimado definido por,

$$cv(\hat{\beta}_k^{*b}) = \frac{1}{\hat{\beta}_k^{*b}} \left(\sum_{i=1}^{n} (\mathbf{x}_{ik}^*)^2 (s^*)^2 \right)^{\frac{1}{2}},$$

onde $(s^*)^2 = \frac{1}{n-1} \sum_{i=1}^{n} (e_i^*)^2$, sendo e^* o resíduo estimado, e $b = 1, 2, \ldots, B_r$, onde B_r significa o número de replicações bootstrap (ver pág. 27).

Calculamos em seguida a média dessas observações, isto é,

$$\overline{cv(\hat{\beta}_k^*)} = \frac{1}{B_r} \sum_{b=1}^{B_r} cv(\hat{\beta}_k^{*b}), k = 0, \ldots, p.$$

O nosso objetivo é verificar o que acontece com os coeficientes de variação de $\hat{\beta}$ relativos às variáveis envolvidas em colinearidade, e esperamos um comportamento diferenciado a medida que o grau de dependência aumenta.

3.3.2 Número e Índices de Condição

Calculamos os índices e o número de condição, equações (2.9) e (2.10), para as replicações bootstrap da matriz $\mathbf{X}$ normalizada. Definimos então a média dos números de condição bootstrap correspondente a cada número de replicação, $\overline{NC_r^*}$, isto é,

$$\overline{NC_r^*} = \frac{1}{B_r} \sum_{b=1}^{B_r} K_b(X^{*t} X^*),$$

onde $K_b(X^t X) = \left(\frac{\lambda_{\max}}{\lambda_{\min}} \right)^{\frac{1}{2}}$ e B_r é dado na pág. 27, $r = 1, \ldots, 7$.

Calculamos em seguida a média dos números de condição entre as replicações bootstrap, definida por

$$\overline{NC} = \frac{1}{7}\sum_{r=1}^{7}\overline{NC_r^*}$$

A média dos índices de condição, $\overline{IC_{rj}^*}$ é definida como:

$$\overline{IC_{rj}^*} = \frac{1}{B_r}\sum_{b=1}^{B_r}\eta_{bj}^*, \quad j = 1, \ldots, p$$

onde $\eta_{bj}^* = \left(\frac{\lambda_{max}}{\lambda_j}\right)^{\frac{1}{2}}$ e onde os índices são os mesmos definidos para o número de condição.

E ainda a média dos índices de condição entre as replicações bootstrap é definida por:

$$\overline{IC_j^*} = \frac{1}{7}\sum_{r=1}^{7}\overline{IC_{rj}^*}, \quad j = 1, \ldots, p$$

3.3.3 Proporção de Decomposição daVariância

Um outro método, também proposto por BKW, é a utilização da proporção de decomposição da variância como diagnóstico da colinearidade, equação (2.17).

Calculamos a proporção de decomposição da variância média, definida por

$$(\bar{\pi}_{jk}^*)_r = \frac{1}{B_r}\sum_{l=1}^{B_r}\pi_{jkl}^*, \quad j, k = 1, 2 \ldots, p$$

onde π_{jkl}^* representa cada proporção de decomposição da variância calculada para cada replicação bootstrap e definida no Capítulo 2.

A seguir calculamos a média entre as replicações bootstrap:

$$\bar{\pi}_{jk}^* = \frac{1}{7}\sum_{r=1}^{7}\bar{\pi}_{jk_r}^*, \quad j, k = 1, \ldots, p$$

3.3.4 As razões $\bar{q}(A_i^*/A_0^*), i = 1, \ldots, p$

O método bootstrap nos propicia a estimação (ou aproximação) da distibuição empírica para cada $\hat{\beta}_j$ $(j = 0, 1, \ldots, 5)$ e, consequentemente seus intervalos de confiança. Consideremos um caso genérico onde uma amostra bootstrap é gerada de acordo com $\hat{P} \rightarrow x^*$, ver equação (3.2) e as replicações bootstrap $\hat{\beta}_b^* = T(\mathbf{X}_b^*)$ são calculadas. Seja $\hat{G}_j^*$ a função de distribuição acumulada de $\hat{\beta}_j^*$. O intervalo de confiança central de nível $1-2\alpha$ baseado nos percentis α e $1-\alpha$ de $\hat{G}_j^*$ é definido por:

$$\left[\hat{\beta}_{j,l}, \hat{\beta}_{j,u}\right] = \left[\hat{G}_j^{*-1}(\alpha), \hat{G}_j^{*-1}(1-\alpha)\right], j = 0, 1, \ldots, 5 \tag{3.3}$$

Como $\hat{G}_j^{*-1}(\alpha) = \hat{\beta}_j^{*(\alpha)}$, o intervalo de confiança baseado no percentil pode ser escrito como:

$$\left[\hat{\beta}_{j,l}, \hat{\beta}_{j,u}\right] = \left[\hat{\beta}_j^{*(\alpha)}, \hat{\beta}_j^{*(1-\alpha)}\right] \tag{3.4}$$

e sua amplitude

$$A_j^* = \hat{\beta}_j^{*(1-\alpha)} - \hat{\beta}_j^{*(\alpha)},$$

A utilização do método percentil para o cálculo do intervalo de confiança está baseada na suposição de que as distribuições $\hat{G}_j^*$, são simétricas, caso isso não ocorra outros métodos para o cálculo dos intervalos de confiança devem ser utilizados, como por exemplo o método do percentil ajustado, ver Efron e Tibshirani (1993)[6].

Através da utilização do método percentil, para o cálculo de intervalos de confiança, derivamos um outro diagnóstico de colinearidade. Escolhemos como nível de significância o valor de 5% para cada coeficiente estimado $\hat{\beta}_j$ $j = 0, \ldots, 5$, em seguida calculamos as amplitudes dos intervalos, de confiança para cada $\hat{\beta}_j^*$ e definimos a razão $\bar{q}(A_i^*/A_0^*)$ como:

$$\bar{q}(A_i^*/A_0^*) = \frac{\sum_{j=1}^{7} A_j^*(\hat{\beta}_i^*)}{\sum_{j=1}^{7} A_j^*(\hat{\beta}_0^*)},\ i = 1, \ldots, 5$$

Não calculamos $\bar{q}$ para $i = 0$, pois obviamente $\bar{q}(A_0^*/A_0^*) = 1$.

Isto é, somamos as amplitudes dos intervalos de confiança $A_j^*(\hat{\beta}_i^*)$ para cada número de replicação bootstrap ($j = 1, 2, \ldots, 7$) e dividimos pela soma das amplitudes dos intervalos de confiança para $\hat{\beta}_0^*$, $A_j^*(\hat{\beta}_0^*)$, gerando assim a razão $\bar{q}(A_i^*/A_0^*)$. A nossa motivação ao derivarmos um diagnóstico baseado no intervalo de confiança para cada $\hat{\beta}_j^*$ vem do fato que as variáveis envolvidas em colinearidade deveriam apresentar intervalos mais largos do que as outras, e que ao calcularmos as amplitudes dos intervalos, e depois a razão entre estas amplitudes e uma amplitude considerada como padrão, relativa ao coeficiente $\hat{\beta}_0^*$, conseguiriamos ter um indicativo das variáveis envolvidas em colinearidade.

Quando nos referirmos a razão $\bar{q}(A_i^*/A_0^*)$, durante o restante do trabalho, e não houver possibilidade de confusão em relação ao índice i, utilizaremos somente $\bar{q}$.

3.4 Resultados

3.4.1 Modelo Básico

Com intuito de avaliar a sensibilidade dos diagnósticos propostos, foram gerados, através de simulações Monte Carlo, nove conjuntos de dados (com sementes de geração diferentes e ímpares), segundo o modelo $\mathbf{y} = \mathbf{X}\beta + \epsilon$, com cinco variáveis $(X_1, X_2, X_3, X_4, X_5)$ e cinquenta observações cada uma, com grau de colinearidade crescente entre as variáveis X_1 e X_2, isto é, $X_0 = [1, \ldots, 1]$ é o termo constante, X_1, X_3, X_4 e X_5 foram gerados tendo distribuição $N(0, 1)$ e X_2 foi gerado como:

$$X_2 = X_1 + f\epsilon_1$$

sendo os fatores f dados por $(3, \frac{3}{2}, 1, \frac{2}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{12}, \frac{1}{24}, \frac{1}{48})$, onde f decrescente implica no crescimento do grau de colineraidade, e $\epsilon_1 \sim N(0,1)$.

Para a construção de **y** supomos $\epsilon \sim N(0,1)$ e definimos os seguintes coeficientes para os $\beta's$:

$$\beta_0 = 4, \beta_1 = 2, \beta_2 = 5, \beta_3 = -2, \beta_4 = 1, \beta_5 = 3 \tag{3.5}$$

Desse modo obtivemos nove modelos com grau crescente de colinearidade entre as variáveis X_1 e X_2. Aplicamos nesses nove modelos os quatro diagnósticos apresentados na secção anterior e avaliamos o desempenho de cada um.

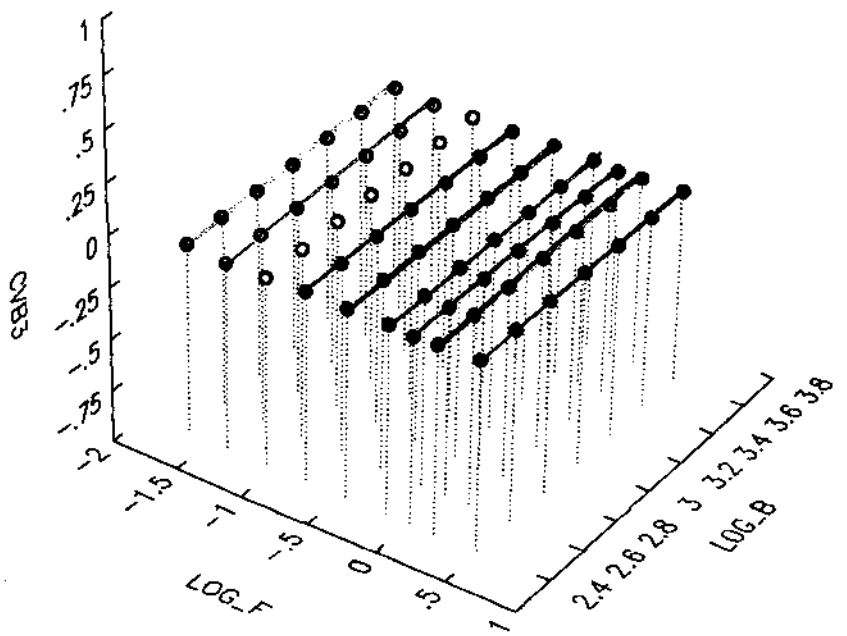
O primeiro diagnóstico avaliado é baseado na instabilidade apresentada pelo coeficiente de variação estimado de $\hat{\beta}_j^*$ na presença de variáveis colineares. Para melhor visualização da diferença de comportamento, do coeficiente de variação estimado de $\hat{\beta}_j^*$, entre as variáveis que estão envolvidas em colineariddae e as que não estão, construimos gráficos envolvendo a média do $cv(\hat{\beta}_i^*)$, o $log(f\ )$ e $log(B_r)$.

Os gráficos mostram, para as variáveis que não estão envolvidas em colinearidade, isto é, para X_3, X_4 e X_5, um patamar próximo de zero. Independente do fator de colinearidade que esteja envolvendo X_1 e X_2. A média do coeficiente de variação estimado de cada $\hat{\beta}_j^*$ se mantém constante, como pode ser visto no Gráfico 1 para o coeficiente de variação de $\hat{\beta}_3^*$. Os gráficos para os coeficientes $\hat{\beta}_4^*$ e $\hat{\beta}_5^*$ são semelhantes.

Gráfico 1

Média do Coeficiente de Variação

Estimado de $\hat{\beta}_3^*$



Entretanto, para as variáveis envolvidas em colinearidade o patamar não se mantém. Pudemos notar que para a média do $cv(\hat{\beta}_1^*)$ a instabilidade causada pela colinearidade já pode ser percebida, mesmo que suavemente, a partir do fator de colinearidade $f = \frac{1}{3}$, tornando-se cada vez mais evidente à medida em que o grau de colinearidade aumenta. Para a média do $cv(\hat{\beta}_2^*)$ a presença da colinearidade demora um pouco mais a aparecer, isto é, no fator de colinearidade $f = \frac{1}{6}$ a instabilidade começa a aparecer suavemente tornando-se cada vez mais visível à medida em que o grau de colinearidade aumenta. Esta avaliação foi feita para a variação do $cv(\hat{\beta}_i^*)$ entre $(-1, 1)$. A variação na instabilidade do coeficiente de variação estimado, à medida em que o grau de colinearidade entre as

variáveis X_1 e X_2, aumenta pode ser visto claramente nos Gráficos 2 e 3.

Gráfico 2

Média do Coeficiente de Variação

Estimado de $\hat{\beta}_1^*$

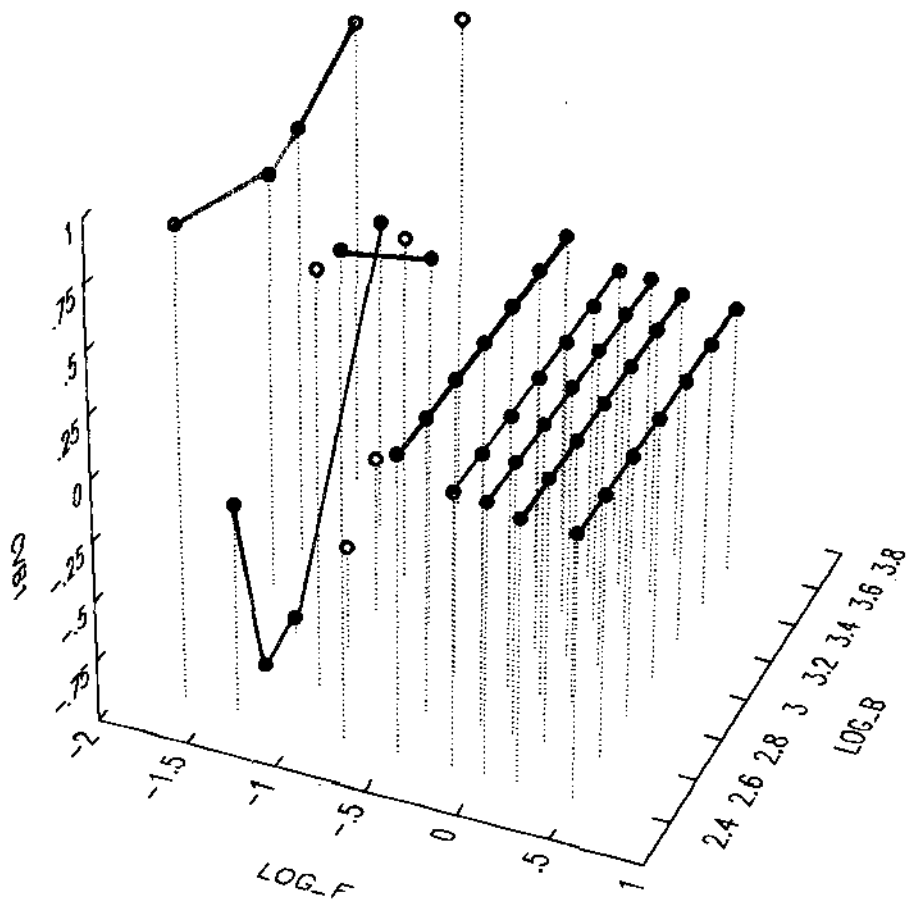
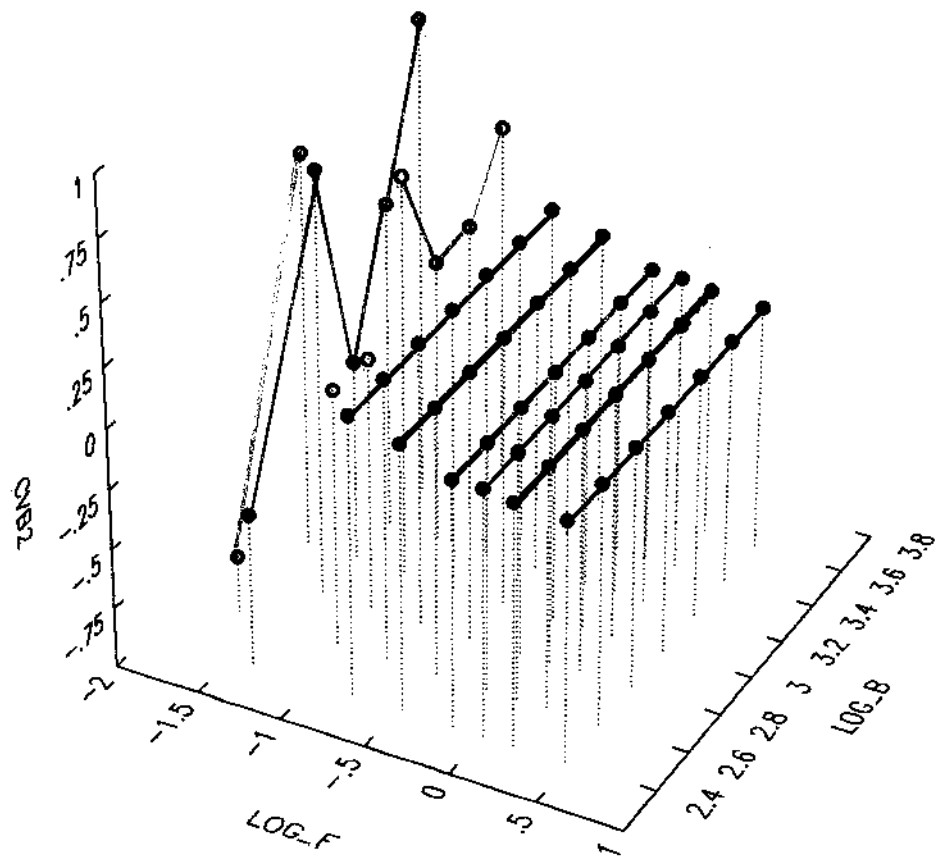


Gráfico 3

Média do Coeficiente de Variação

Estimado de $\hat{\beta}_2^*$



Para avaliar o diagnóstico proposto por BKW, calculamos os índices de condição médio, sendo que o número de condição é igual ao maior índice de condição, como já foi dito no Capítulo 2. Nota-se que, à medida em que o grau de colinearidade aumenta, o valor de $\overline{NC_f^*}$ também aumenta, porém seguindo a orientação de BKW, uma fraca presença da colinearidade seria detectada a partir de $f = \frac{1}{3}$, e uma clara presença da colinearidade só seria detectada a partir do sétimo fator, $f = \frac{1}{12}$, pois somente a partir desse modelo o número de condição médio, $\overline{NC^*}$, é maior do que 30 e está associado a proporção de decomposição da variância média, $\bar{\pi}_{jk}^*$ maior do que 0.5. É importante ressaltar que através do índice de condição temos somente uma indicação da existência

da colinearidade, sendo indispensável a utilização da PDV para descobrir quais são as variáveis envolvidas. Na tabela 3.1 estão os índices de condição médio, $\overline{IC_j^*}$, $j = 1, \ldots, p$. Devemos deixar claro que o índice j $(j = 1, 2, \ldots, p)$ do índice de condição, não está relacionado com o índice de $\hat{\beta}_j^*$ que também varia de 1 até p, mas sim apresentam-se ordenados pela ordem de grandeza dos valores singulares.

Tabela 3.1

$f = 3$	$f = 3/2$	$f = 1$
$\overline{IC_1^*}$=1.1687	$\overline{IC_1^*}$=1.1714	$\overline{IC_1^*}$=1.2552
$\overline{IC_2^*}$=1.2045	$\overline{IC_2^*}$=1.3332	$\overline{IC_2^*}$=1.4215
$\overline{IC_3^*}$=1.5863	$\overline{IC_3^*}$=1.7760	$\overline{IC_3^*}$=1.7188
$\overline{IC_4^*}$=1.8851	$\overline{IC_4^*}$=1.8492	$\overline{IC_4^*}$=2.0498
$\overline{IC_5^*}$=2.3125	$\overline{IC_5^*}$=2.4474	$\overline{IC_5^*}$=3.2633

$f = 2/3$	$f = 1/3$	$f = 1/6$
$\overline{IC_1^*}$=1.2695	$\overline{IC_1^*}$=1.3097	$\overline{IC_1^*}$=1.2669
$\overline{IC_2^*}$=1.5291	$\overline{IC_2^*}$=1.4849	$\overline{IC_2^*}$=1.4789
$\overline{IC_3^*}$=1.8148	$\overline{IC_3^*}$=1.6850	$\overline{IC_3^*}$=1.9739
$\overline{IC_4^*}$=2.1667	$\overline{IC_4^*}$=1.9893	$\overline{IC_4^*}$=2.0766
$\overline{IC_5^*}$=3.6186	$\overline{IC_5^*}$=8.2855	$\overline{IC_5^*}$=12.904

$f = 1/12$	$f = 1/24$	$f = 1/48$
$\overline{IC_1^*}$=1.2734	$\overline{IC_1^*}$=1.2424	$\overline{IC_1^*}$=1.3016
$\overline{IC_2^*}$=1.5047	$\overline{IC_2^*}$=1.4578	$\overline{IC_2^*}$=1.5061
$\overline{IC_3^*}$=1.7487	$\overline{IC_3^*}$=1.7197	$\overline{IC_3^*}$=1.7611
$\overline{IC_4^*}$=2.1637	$\overline{IC_4^*}$=2.0504	$\overline{IC_4^*}$=2.2201
$\overline{IC_5^*}$=35.012	$\overline{IC_5^*}$=67.744	$\overline{IC_5^*}$=132.91

E a proporção de decomposição da variância é vista na tabela 3.2, a partir do fator de colinearidade $f = \frac{1}{3}$.

Tabela 3.2

Fator	Índice	$\bar{\pi}^*_{15}$	$\bar{\pi}^*_{25}$	$\bar{\pi}^*_{35}$	$\bar{\pi}^*_{45}$	$\bar{\pi}^*_{55}$
$f = \frac{1}{3}$	$\overline{IC^*_5}$	0.9771	0.9798	0.1283	0.0522	0.0863
$f = \frac{1}{6}$	$\overline{IC^*_5}$	0.9918	0.9914	0.0424	0.0585	0.0249
$f = \frac{1}{12}$	$\overline{IC^*_5}$	0.9988	0.9988	0.0282	0.0241	0.0865
$f = \frac{1}{24}$	$\overline{IC^*_5}$	0.9997	0.9997	0.1003	0.0354	0.0560
$f = \frac{1}{48}$	$\overline{IC^*_5}$	0.9997	0.9998	0.0203	0.0180	0.0541

Finalmente, analisando conjuntamente os quatro diagnósticos, achamos que a razão $\bar{q}$ mostrou-se mais eficiente na detecção da presença de colinearidade.

Através desse procedimento pudemos verificar que o valor de $\bar{q}(A^*_2/A^*_0)$, já no primeiro fator de colinearidade, $f = 3$, na tabela 3.3, aparece como aproximadamente o inverso do fator de colinearidade se destacando dos demais, que estão próximos de 1. Este é um fato importante e poderia ser um indicativo de que a variável X_2 está envolvida em alguma relação de dependência, mas até esse momento não sabemos com qual variável. À medida em que o grau de colinearidade aumenta, isto é, a partir de $f = \frac{1}{3}$, percebemos que a razão $\bar{q}(A^*_1/A^*_0)$ também se destaca das demais ficando cada vez mais evidente a diferença das razões $\bar{q}(A^*_i/A^*_0)$ para $i = 1$ e 2, em relação as demais variáveis, sendo sempre aproximadamente o inverso do fator de colinearidade entre as variáveis X_1 e X_2. O valor das razões $\bar{q}(A^*_i/A^*_0)$ para $i = 3, 4$ e 5 continua sempre próximo de um, independente do grau de colinearidade entre as variáveis, indicando, desse modo, que estas variáveis não estão envolvidas em colinearidade.

Tabela 3.3

Modelo supondo colinearidade crescente entre X_1 e X_2, sendo X_1, X_3, X_4 e $X_5 \sim N(0,1)$, $\epsilon \sim N(0,1)$ e $\epsilon_1 \sim N(0,1)$.

$f = 3$
$\bar{q}(A_1^*/A_0^*) = 0.8475$
$\bar{q}(A_2^*/A_0^*) = 0.3155$
$\bar{q}(A_3^*/A_0^*) = 0.9001$
$\bar{q}(A_4^*/A_0^*) = 1.0969$
$\bar{q}(A_5^*/A_0^*) = 1.3193$

$f = 3/2$
$\bar{q}(A_1^*/A_0^*) = 0.9890$
$\bar{q}(A_2^*/A_0^*) = 0.5128$
$\bar{q}(A_3^*/A_0^*) = 0.9153$
$\bar{q}(A_4^*/A_0^*) = 0.9598$
$\bar{q}(A_5^*/A_0^*) = 1.4501$

$f = 1$
$\bar{q}(A_1^*/A_0^*) = 1.6498$
$\bar{q}(A_2^*/A_0^*) = 1.1279$
$\bar{q}(A_3^*/A_0^*) = 1.2604$
$\bar{q}(A_4^*/A_0^*) = 1.0234$
$\bar{q}(A_5^*/A_0^*) = 0.9969$

$f = 2/3$
$\bar{q}(A_1^*/A_0^*) = 1.7228$
$\bar{q}(A_2^*/A_0^*) = 1.4592$
$\bar{q}(A_3^*/A_0^*) = 1.4842$
$\bar{q}(A_4^*/A_0^*) = 1.2139$
$\bar{q}(A_5^*/A_0^*) = 1.1257$

$f = 1/3$
$\bar{q}(A_1^*/A_0^*) = 3.5616$
$\bar{q}(A_2^*/A_0^*) = 2.9846$
$\bar{q}(A_3^*/A_0^*) = 1.1024$
$\bar{q}(A_4^*/A_0^*) = 1.3473$
$\bar{q}(A_5^*/A_0^*) = 0.9966$

$f = 1/6$
$\bar{q}(A_1^*/A_0^*) = 5.0150$
$\bar{q}(A_2^*/A_0^*) = 4.7590$
$\bar{q}(A_3^*/A_0^*) = 1.0882$
$\bar{q}(A_4^*/A_0^*) = 1.0180$
$\bar{q}(A_5^*/A_0^*) = 1.2725$

$f = 1/12$
$\bar{q}(A_1^*/A_0^*) = 15.723$
$\bar{q}(A_2^*/A_0^*) = 15.908$
$\bar{q}(A_3^*/A_0^*) = 1.1676$
$\bar{q}(A_4^*/A_0^*) = 0.9218$
$\bar{q}(A_5^*/A_0^*) = 1.3966$

$f = 1/24$
$\bar{q}(A_1^*/A_0^*) = 30.438$
$\bar{q}(A_2^*/A_0^*) = 30.443$
$\bar{q}(A_3^*/A_0^*) = 1.1157$
$\bar{q}(A_4^*/A_0^*) = 0.8855$
$\bar{q}(A_5^*/A_0^*) = 1.3242$

$f = 1/48$
$\bar{q}(A_1^*/A_0^*) = 42.057$
$\bar{q}(A_2^*/A_0^*) = 42.291$
$\bar{q}(A_3^*/A_0^*) = 1.0347$
$\bar{q}(A_4^*/A_0^*) = 1.0858$
$\bar{q}(A_5^*/A_0^*) = 0.9235$

Outro fato interessante a ressaltar é que $\bar{q}(A_i^*/A_0^*)$, $i = 1, 2$ pode ser escrito como uma função do fator de colinearidade através da equação de regressão,

$$\log\ q = \alpha - \gamma \log(f)$$

resultando em

$$q = 10^{\alpha} f^{-\gamma} \tag{3.6}$$

Neste caso temos para $\bar{q}(A_1^*/A_0^*)$ que $\hat{\alpha} = 0.1908$ e $\hat{\gamma}$= 0.8569, fazendo com que a equação (3.6) seja dada, numericamente, por:

$$\bar{q} = 1.5517 \times f^{-0.8569}.$$

Os valores estimados e seus respectivos erros relativos, dados por

$$\mid \Delta\% \mid = \frac{\mid Obs - Est \mid}{Obs} \times 100, \tag{3.7}$$

podem ser vistos na tabela abaixo:

Tabela 3.4

Fator	Obs.	Est.	$\mid \Delta\% \mid$
3	0.8475	0.6053	28.58
$\frac{3}{2}$	0.9890	1.0963	10.85
1	1.6498	1.5517	5.946
$\frac{2}{3}$	1.7228	2.1963	27.48
$\frac{1}{3}$	3.5616	3.9779	16.69
$\frac{1}{6}$	5.0150	7.2045	43.66
$\frac{1}{12}$	15.7228	13.0484	17.01
$\frac{1}{24}$	30.4376	23.6326	22.36
$\frac{1}{48}$	42.0568	42.8019	1.772

e a média dos erros é:

$$\frac{1}{9}\sum \mid \Delta\% \mid = 19.37$$

Para $\bar{q}(A_2^*/A_0^*)$ temos que $\hat{\alpha} = -0,0220$ e $\hat{\gamma} = 1,0343$, fazendo com que a equação (3.6) seja dada, numericamente, por:

$$\bar{q} = 0,9505 \times f^{-1,0343}$$

Os valores estimados pela equação de regressão juntamente com os seus erros relativos podem ser vistos na tabela abaixo:

Tabela 3.5

Fator	Obs.	Est.	$\mid \Delta\% \mid$
3	0.3155	0.3051	3.30
$\frac{3}{2}$	0.5128	0.6249	21.86
1	1.1279	0.9505	15.73
$\frac{2}{3}$	1.4592	1.4457	0.9253
$\frac{1}{3}$	2.9846	2.9610	0.7907
$\frac{1}{6}$	4.7590	6.0645	27.43
$\frac{1}{12}$	15.9082	12.4208	21.92
$\frac{1}{24}$	30.4434	25.4393	16.44
$\frac{1}{48}$	42.2909	52.1027	23.20

e a média dos erros é:

$$\frac{1}{9}\sum \mid \Delta\% \mid = 14.62 \tag{3.8}$$

Avaliando essas duas equações de regressão percebemos que a razão $\bar{q}(A_2^*/A_0^*)$ é aproximadamente o inverso do fator de colinearidade. O mesmo não acontecendo exatamente para a razão $\bar{q}(A_1^*/A_0^*)$ pois, como foi visto na Tabela 3.3, $\bar{q}(A_1^*/A_0^*)$ demora mais para manifestar a presença da colinearidade, afetando a estimação da equação de regressão.

O último diagnóstico mostrou-se bastante eficiente em detectar a colinearidade, pois além de indicar a sua presença, mostra ainda quais são as variáveis envolvidas, e o mais importante, dá uma indicação do grau de colinearidade entre as variáveis. A maior vantagem em relação ao coeficiente de variação estimado é que a presença da colinearidade pode ser percebida praticamente ao mesmo tempo em X_1 e X_2. Isto não acontece utilizando o coeficiente de variação estimado, apesar da detecção da presença da colinearidade ser feita nos primeiros fatores para a variável X_1, ela só é percebida para a variável X_2 quando o grau de colinearidade entre as variáveis é severo.

Em relação ao diagnóstico proposto por BKW, podemos observar que ao seguirmos a sua orientação, teriamos um indicativo da presença da colinearidade somente a partir do fator $f = \frac{1}{6}$ e ainda precisariamos da PDV para identificar quais as variáveis envolvidas em colinearidade. Neste exemplo, a razão $\bar{q}$ foi capaz de identificar o grau de colinearidade aproximado e também quais as variáveis envolvidas em colinearidade, sem a necessidade de nenhum outro diagnóstico, isto é, a razão $\bar{q}$ conseguiu mais do que os outros diagnósticos, em poder de detecção.

Como este foi o diagnóstico mais sensível entre os apresentados, vamos a partir de agora verificar o seu comportamento em algumas situações.

Nas duas próximas seções serão apresentados os resultados dos modelos simulados. O leitor irá se deparar com um grande número de tabelas que resumem estes resultados. Apesar de parecer um pouco extensa, a apresentação dessas tabelas é muito importante, já que todo o trabalho foi baseado em resultados de simulações Monte Carlo, portanto experimentais. Como ainda não temos domínio da teoria que fundamenta nossas suposições, essa apresentação se torna indispensável para que se tenha um idéia sequencial e comparativa do potencial do diagnóstico proposto, $\bar{q}$, nos modelos estudados.

As tabelas dos resultados estão divididas em 5 blocos. O primeiro bloco mostra os resultados obtidos com as alterações feitas somente no desvio-padrão de ϵ_1 (tabelas 3.6 à 3.10). O segundo bloco mostra as alterações feitas somente no desvio-padrão das variáveis X_i (tabelas 3.11 à 3.15). O terceiro bloco mostra as duas alterações anteriores combinadas simultaneamente (tabelas 3.16 à 3.23). O quarto bloco mostra os modelos onde as distribuições de X_i são $N(0, \sigma_i^2)$, $\sigma_i^2 \neq 1$, $i = 1, 2, \ldots, 5$ (tabelas 3.24 à 3.27). Finalmente o quinto bloco mostra os modelos com três variáveis colineares (tabelas 3.28 à 3.37).

3.4.2 Modelos com Duas Variáveis Colineares

Modelo de Referência

Já que entre os quatro diagnósticos considerados, a razão $\bar{q}$ foi a mais eficiente em detectar a presença de variáveis colineares, vamos agora verificar o seu comportamento em uma situação ideal, isto é, onde nenhuma das variáveis está envolvida em colinearidade, e vamos considerar este modelo como o de referência.

O modelo de referência foi gerado com 50 observações supondo X_1, X_2, X_3, X_4 e X_5 variáveis aleatórias independentes com distribuição Normal, com média zero e variância um. X_0 é o vetor de constantes iguais a 1. Utilizamos também, a distribuição $N(0,1)$ para para a gerar ϵ e para a construção do vetor de respostas **y** os coeficientes foram determinados como no modelo básico (3.5).

Verificamos que nessa situação a razão $\bar{q}$, mantém-se próxima de 1 para todas as variáveis, como pode ser visto na tabela 3.6. Analisando a tabela percebemos que os valores de $\bar{q}(A_i^*/A_0^*)$ para $i = 3$ e 5 estão um pouco mais distantes de 1 do que os demais. Talvez este fato possa estar relacionado com as sementes utilizadas nas simulações, pois apesar de serem números grandes e ímpares, considerados adequados como sementes de um gerador congruencial multiplicativo de módulo 2, não fizemos durante o nosso trabalho, nenhuma avaliação de alguma possível influência das sementes escolhidas nos resultados obtidos.

Tabela 3.6

Modelo sem colinearidade

supondo X_1, X_2, X_3, X_4 e $X_5 \sim N(0,1)$, $\epsilon \sim N(0,1)$

referência
$\bar{q}(A_1^*/A_0^*) = 1.1856$
$\bar{q}(A_2^*/A_0^*) = 1.0920$
$\bar{q}(A_3^*/A_0^*) = 0.8418$
$\bar{q}(A_4^*/A_0^*) = 1.0647$
$\bar{q}(A_5^*/A_0^*) = 0.7995$

Com o modelo de referência estabelecemos para a razão $\bar{q}$ um patamar próximo de 1, para as variáveis que não estão envolvidas em colinearidade. Entretanto, achamos que seria interessante verificar o que aconteceria com a razão $\bar{q}$ e consequentemente com o patamar estabelecido pelo modelo básico e o de referência, se introduzissemos algumas variações no desvio-padrão de ϵ_1. Sabemos que ϵ_1 está diretamente ligado a construção da variável X_2, pois $X_2 = X_1 + f\epsilon_1$, e até o momento consideramos a distribuição de ϵ e ϵ_1 como $N(0,1)$.

3.4.2.1-Alterações no Desvio de ϵ_1 no Modelo Básico

Analisando as tabelas 3.7 e 3.8 percebemos que os valores de $\bar{q}(A_1^*/A_0^*)$ e $\bar{q}(A_2^*/A_0^*)$ se alteram com as modificações feitas no desvio-padrão de ϵ_1, sendo que para $\bar{q}(A_2^*/A_0^*)$ essa modificação é bem evidente, pois se compararmos os valores de $\bar{q}(A_2^*/A_0^*)$ nos 3 modelos, tabelas 3.3, 3.7 e 3.8, vemos que quando o desvio-padrão de ϵ_1 é igual a $\frac{1}{2}$ o valor da razão $\bar{q}(A_2^*/A_0^*)$, que era igual a 0.3155 supondo desvio de $\epsilon_1 = 1$, passa a ser 0.6160. E quando o desvio de ϵ_1 é igual a 2 a razão $\bar{q}(A_2^*/A_0^*)$ passa a ser 0.1578. Para $\bar{q}(A_1^*/A_0^*)$ essa alteração não é tão clara nos primeiros fatores de colinearidade, ficando mais visível

a medida que o grau de colinearidade aumenta, isto é, a partir de $f = \frac{2}{3}$. Podemos fazer uma tentativa de reescrever a equação (3.6) levando em consideração as variações ocorridas em $\bar{q}(A_i^*/A_0^*)$ para $i = 1$ e 2, através da seguinte equação de regressão,

$$\bar{q} = 10^{\alpha} \left(\sigma_{\epsilon_1} \times f\right)^{-\gamma} \tag{3.9}$$

Por outro lado, nada acontece com a razão $\bar{q}(A_i^*/A_0^*)$ para $i = 3, 4$ e 5, isto é, nas duas variações feitas em ϵ_1 o patamar de valores próximos de 1 se mantém, como já tinhamos visto na tabela 3.3, onde $\epsilon_1 \sim N(0,1)$.

Tabela 3.7

Modelo supondo colinearidade crescente entre X_1 e X_2,
sendo X_1, X_3, X_4 e $X_5 \sim N(0,1)$, $\epsilon \sim N(0,1)$ e $\epsilon_1 \sim N(0, \frac{1}{2}^2)$.

$f = 3$	$f = 3/2$	$f = 1$
$\bar{q}(A_1^*/A_0^*) = 1.0823$	$\bar{q}(A_1^*/A_0^*) = 1.2298$	$\bar{q}(A_1^*/A_0^*) = 2.6142$
$\bar{q}(A_2^*/A_0^*) = 0.6160$	$\bar{q}(A_2^*/A_0^*) = 1.0255$	$\bar{q}(A_2^*/A_0^*) = 2.2556$
$\bar{q}(A_3^*/A_0^*) = 0.8787$	$\bar{q}(A_3^*/A_0^*) = 0.9153$	$\bar{q}(A_3^*/A_0^*) = 1.2603$
$\bar{q}(A_4^*/A_0^*) = 1.0708$	$\bar{q}(A_4^*/A_0^*) = 0.9598$	$\bar{q}(A_4^*/A_0^*) = 1.0233$
$\bar{q}(A_5^*/A_0^*) = 1.2880$	$\bar{q}(A_5^*/A_0^*) = 1.4501$	$\bar{q}(A_5^*/A_0^*) = 0.9968$

$f = 2/3$	$f = 1/3$	$f = 1/6$
$\bar{q}(A_1^*/A_0^*) = 2.9278$	$\bar{q}(A_1^*/A_0^*) = 6.4904$	$\bar{q}(A_1^*/A_0^*) = 9.7309$
$\bar{q}(A_2^*/A_0^*) = 2.9185$	$\bar{q}(A_2^*/A_0^*) = 5.9693$	$\bar{q}(A_2^*/A_0^*) = 9.5181$
$\bar{q}(A_3^*/A_0^*) = 1.48844$	$\bar{q}(A_3^*/A_0^*) = 1.1024$	$\bar{q}(A_3^*/A_0^*) = 1.0882$
$\bar{q}(A_4^*/A_0^*) = 1.2198$	$\bar{q}(A_4^*/A_0^*) = 1.3473$	$\bar{q}(A_4^*/A_0^*) = 1.0180$
$\bar{q}(A_5^*/A_0^*) = 1.1257$	$\bar{q}(A_5^*/A_0^*) = 0.9966$	$\bar{q}(A_5^*/A_0^*) = 1.2725$

A tabela continua na página seguinte

$f = 1/12$	$f = 1/24$	$f = 1/48$
$\bar{q}(A_1^*/A_0^*) = 31.459$	$\bar{q}(A_1^*/A_0^*) = 60.894$	$\bar{q}(A_1^*/A_0^*) = 84.185$
$\bar{q}(A_2^*/A_0^*) = 31.816$	$\bar{q}(A_2^*/A_0^*) = 60.887$	$\bar{q}(A_2^*/A_0^*) = 84.582$
$\bar{q}(A_3^*/A_0^*) = 1.1676$	$\bar{q}(A_3^*/A_0^*) = 1.1157$	$\bar{q}(A_3^*/A_0^*) = 1.0347$
$\bar{q}(A_4^*/A_0^*) = 0.9218$	$\bar{q}(A_4^*/A_0^*) = 0.8855$	$\bar{q}(A_4^*/A_0^*) = 1.1019$
$\bar{q}(A_5^*/A_0^*) = 1.3966$	$\bar{q}(A_5^*/A_0^*) = 1.3242$	$\bar{q}(A_5^*/A_0^*) = 0.9235$

Neste caso, onde $\sigma_{\epsilon_1} = \frac{1}{2}$, temos $\hat{\alpha} = 0.0801$ e $\hat{\gamma}= 0.9563$ para $\bar{q}(A_1^*/A_0^*)$, fazendo com que a equação (3.9) seja dada, numericamente, por:

$$\bar{q} = 1.2025 \times (\frac{1}{2} \times f)^{-0.9563}$$

e para $\bar{q}(A_2^*/A_0^*)$ temos $\hat{\alpha} = -0.0356$ e $\hat{\gamma}= 1.0368$, fazendo com que a equação (3.9) seja dada, numericamente, por:

$$\bar{q} = 0.9213 \times (\frac{1}{2} \times f)^{-1.0368}.$$

Tabela 3.8

Modelo supondo colinearidade crescente entre X_1 e X_2,
sendo X_1, X_3, X_4 e $X_5 \sim N(0,1)$, $\epsilon \sim N(0,1)$ e $\epsilon_1 \sim N(0, 2^2)$.

$f = 3$	$f = 3/2$	$f = 1$
$\bar{q}(A_1^*/A_0^*) = 0.8666$	$\bar{q}(A_1^*/A_0^*) = 0.9603$	$\bar{q}(A_1^*/A_0^*) = 1.2514$
$\bar{q}(A_2^*/A_0^*) = 0.1578$	$\bar{q}(A_2^*/A_0^*) = 0.2564$	$\bar{q}(A_2^*/A_0^*) = 0.5639$
$\bar{q}(A_3^*/A_0^*) = 0.9001$	$\bar{q}(A_3^*/A_0^*) = 0.9153$	$\bar{q}(A_3^*/A_0^*) = 1.2603$
$\bar{q}(A_4^*/A_0^*) = 1.0969$	$\bar{q}(A_4^*/A_0^*) = 0.9598$	$\bar{q}(A_4^*/A_0^*) = 1.0233$
$\bar{q}(A_5^*/A_0^*) = 1.3193$	$\bar{q}(A_5^*/A_0^*) = 1.4501$	$\bar{q}(A_5^*/A_0^*) = 0.9968$

$f = 2/3$	$f = 1/3$	$f = 1/6$
$\bar{q}(A_1^*/A_0^*) = 1.3298$	$\bar{q}(A_1^*/A_0^*) = 2.2064$	$\bar{q}(A_1^*/A_0^*) = 2.7383$
$\bar{q}(A_2^*/A_0^*) = 0.7296$	$\bar{q}(A_2^*/A_0^*) = 1.4923$	$\bar{q}(A_2^*/A_0^*) = 2.3795$
$\bar{q}(A_3^*/A_0^*) = 1.4884$	$\bar{q}(A_3^*/A_0^*) = 1.1024$	$\bar{q}(A_3^*/A_0^*) = 1.0882$
$\bar{q}(A_4^*/A_0^*) = 1.2198$	$\bar{q}(A_4^*/A_0^*) = 1.3473$	$\bar{q}(A_4^*/A_0^*) = 1.0180$
$\bar{q}(A_5^*/A_0^*) = 1.1257$	$\bar{q}(A_5^*/A_0^*) = 0.9966$	$\bar{q}(A_5^*/A_0^*) = 1.2725$

$f = 1/12$	$f = 1/24$	$f = 1/48$
$\bar{q}(A_1^*/A_0^*) = 7.8046$	$\bar{q}(A_1^*/A_0^*) = 15.255$	$\bar{q}(A_1^*/A_0^*) = 20.954$
$\bar{q}(A_2^*/A_0^*) = 7.9541$	$\bar{q}(A_2^*/A_0^*) = 15.222$	$\bar{q}(A_2^*/A_0^*) = 21.145$
$\bar{q}(A_3^*/A_0^*) = 1.1676$	$\bar{q}(A_3^*/A_0^*) = 1.1157$	$\bar{q}(A_3^*/A_0^*) = 1.0347$
$\bar{q}(A_4^*/A_0^*) = 0.9218$	$\bar{q}(A_4^*/A_0^*) = 0.8855$	$\bar{q}(A_4^*/A_0^*) = 1.1019$
$\bar{q}(A_5^*/A_0^*) = 1.3966$	$\bar{q}(A_5^*/A_0^*) = 1.3242$	$\bar{q}(A_5^*/A_0^*) = 0.9235$

Neste caso, onde $\sigma_{\epsilon_1} = 2$, temos $\hat{\alpha} = 0.3089$ e $\hat{\gamma}$= 0.6965 para $\bar{q}(A_1^*/A_0^*)$, fazendo com que a equação (3.9) seja dada, numericamente, por:

$$\bar{q} = 2.0365 \times (2 \times f)^{-0.6965}$$

e para $\bar{q}(A_2^*/A_0^*)$ temos $\hat{\alpha} = -0.0117$ e $\hat{\gamma}$= 1.0342, fazendo com que a equação (3.9) seja dada, numericamente, por:

$$\bar{q} = 0.9734 \times (2 \times f)^{-1.0342}$$

Com as modificações feitas no desvio-padrão de ϵ_1, observamos que os valores de 10^{α} e γ se alteram para a razão $\bar{q}(A_1^*/A_0^*)$, o mesmo não acontecendo para $\bar{q}(A_2^*/A_0^*)$ como pode ser visto nas tabelas 3.9 e 3.10. Suspeitamos que isto ocorra pois o peso da colinearidade é maior em X_1, que é a variável geradora de X_2. Outro fato observado está relacionado com σ_{ϵ_1} e σ_{X_i} (o desvio-padrão da variáveis), percebemos que quando $\sigma_{\epsilon_1} > \sigma_{X_i}$ os valores de 10^{α} e γ sofrem as maiores alterações, sendo neste caso $\hat{\gamma} = -0.6965$, o valor mais distante de um, que está associado ao maior valor de σ_{ϵ_1} em relação ao σ_{X_i}. Suspeitamos que quando $\sigma_{\epsilon_1} > \sigma_{X_i}$ o desvio-padrão da variável X_1 fica mascarado, influenciando no ajuste da equação e levando a uma estimativa mais baixa de $\hat{\gamma}$. Vamos observar nos demais modelos se o fato ocorrido aqui também pode ser verificado.

Tabela 3.9

$\bar{\mathbf{q}}(\mathbf{A_1^*/A_0^*})$

σ_{X_i}	σ_{ϵ_1}	10^{α}	γ
1	$\frac{1}{2}$	1.2025	-0.9563
1	1	1.5517	-0.8569
1	2	2.0365	-0.6965

e

Tabela 3.10

$\bar{\mathbf{q}}(\mathbf{A_2^*/A_0^*})$

σ_{X_i}	σ_{ϵ_1}	10^{α}	γ
1	$\frac{1}{2}$	0.9213	-1.0368
1	1	0.9505	-1.0343
1	2	0.9734	-1.0342

Através dos modelos apresentados vimos que o desvio de ϵ_1 tem influência no valor da razão $\bar{q}(A_i^*/A_0^*)$ para as variáveis envolvidas em colinearidade, no caso para $i = 1$ e 2. Até o momento consideramos que as variáveis X_1, X_3, X_4 e X_5 tinham distribuição $N(0,1)$, e a primeira modificação feita no modelo básico alterou o desvio-padrão de ϵ_1. Decidimos, então explorar mais um ponto e optamos por introduzir algumas modificações no desvio-padrão das variáveis X_1, X_3, X_4 e X_5.

A princípio consideramos um modelo sem colinearidade, onde apenas a variável X_2 tem distribuição $N(0, f^2)$, sendo $f \in \left(3, \frac{3}{2}, 1, \frac{2}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{12}, \frac{1}{24}, \frac{1}{48}\right)$, e as variáveis X_1, X_3, X_4 e $X_5 \sim N(0,1)$. Supomos ainda que $\epsilon \sim N(0,1)$ e para construção de $\mathbf{y}$ os coeficientes foram determinados novamente como no modelo básico.

A construção desse modelo sem colinearidade e com a variável X_2 tendo como desvio-padrão o mesmo valor do fator de colinearidade f, teve como objetivo nos dar uma idéia do que aconteceria com a razão $\bar{q}$ nesta situação, para que nos modelos seguintes, onde vamos combinar a presença da colinearidade e desvio-padrão diferente de 1 para as variáveis X_i, tivessemos maior segurança para separar o efeito de cada modificação na razão $\bar{q}$.

Tabela 3.11

Modelo sem colinearidade supondo

$X_2 \sim N(0, f^2)$, X_1, X_3, X_4 e $X_5 \sim N(0,1)$ e $\epsilon \sim N(0,1)$.

$f = 3$	$f = 3/2$	$f = 1$
$\bar{q}(A_1^*/A_0^*) = 0.7706$	$\bar{q}(A_1^*/A_0^*) = 1.0063$	$\bar{q}(A_1^*/A_0^*) = 0.7921$
$\bar{q}(A_2^*/A_0^*) = 0.3472$	$\bar{q}(A_2^*/A_0^*) = 0.5620$	$\bar{q}(A_2^*/A_0^*) = 1.1504$
$\bar{q}(A_3^*/A_0^*) = 1.0008$	$\bar{q}(A_3^*/A_0^*) = 0.9223$	$\bar{q}(A_3^*/A_0^*) = 0.8413$
$\bar{q}(A_4^*/A_0^*) = 0.9305$	$\bar{q}(A_4^*/A_0^*) = 0.9993$	$\bar{q}(A_4^*/A_0^*) = 0.9948$
$\bar{q}(A_5^*/A_0^*) = 1.1444$	$\bar{q}(A_5^*/A_0^*) = 1.2254$	$\bar{q}(A_5^*/A_0^*) = 1.0275$

$f = 2/3$	$f = 1/3$	$f = 1/6$
$\bar{q}(A_1^*/A_0^*) = 0.7929$	$\bar{q}(A_1^*/A_0^*) = 1.2760$	$\bar{q}(A_1^*/A_0^*) = 0.8621$
$\bar{q}(A_2^*/A_0^*) = 1.8157$	$\bar{q}(A_2^*/A_0^*) = 4.1915$	$\bar{q}(A_2^*/A_0^*) = 6.2064$
$\bar{q}(A_3^*/A_0^*) = 1.2166$	$\bar{q}(A_3^*/A_0^*) = 0.9249$	$\bar{q}(A_3^*/A_0^*) = 1.0096$
$\bar{q}(A_4^*/A_0^*) = 1.0190$	$\bar{q}(A_4^*/A_0^*) = 1.2519$	$\bar{q}(A_4^*/A_0^*) = 0.7062$
$\bar{q}(A_5^*/A_0^*) = 0.6449$	$\bar{q}(A_5^*/A_0^*) = 1.2144$	$\bar{q}(A_5^*/A_0^*) = 0.7030$

$f = 1/12$	$f = 1/24$	$f = 1/48$
$\bar{q}(A_1^*/A_0^*) = 1.1832$	$\bar{q}(A_1^*/A_0^*) = 0.9162$	$\bar{q}(A_1^*/A_0^*) = 0.6804$
$\bar{q}(A_2^*/A_0^*) = 8.9809$	$\bar{q}(A_2^*/A_0^*) = 23.101$	$\bar{q}(A_2^*/A_0^*) = 30.488$
$\bar{q}(A_3^*/A_0^*) = 0.8984$	$\bar{q}(A_3^*/A_0^*) = 1.0390$	$\bar{q}(A_3^*/A_0^*) = 0.8667$
$\bar{q}(A_4^*/A_0^*) = 0.9522$	$\bar{q}(A_4^*/A_0^*) = 0.9841$	$\bar{q}(A_4^*/A_0^*) = 0.5956$
$\bar{q}(A_5^*/A_0^*) = 0.7754$	$\bar{q}(A_5^*/A_0^*) = 1.2730$	$\bar{q}(A_5^*/A_0^*) = 0.8803$

Analisando a tabela 3.11 verificamos que a variável X_2, que apresenta desvio-padrão igual a f , mostra no primeiro fator, $f = 3$, o valor $\bar{q}(A_2^*/A_0^*)$ aproximadamente igual ao inverso do desvio-padrão de X_2. Entretanto, esse fato é manifestado claramente através da razão $\bar{q}$ a partir do fator $f = \frac{1}{3}$, enquanto as demais variáveis mantém o valor de $\bar{q}$ próximo de um, $\bar{q}(A_2^*/A_0^*)$ é sempre, aproximadamente, o inverso do desvio da variável X_2.

Vimos que o fato do desvio-padrão da variável X_2 ser diferente de 1 tem influência sobre o valor da razão $\bar{q}$. Agora vamos analisar modelos onde os desvios-padrão são diferentes de 1 para as variáveis X_1, X_3, X_4 e X_5, mantendo $\epsilon_1 \sim N(0,1)$, e as variáveis X_1 e X_2 tem grau crescente de colinearidade, como no modelo básico.

3.4.2.2-Alterações no Desvio das Variáveis X_1, X_3, X_4 e X_5 mantendo $\epsilon_1 \sim N(0,1)$

Analisando as tabelas 3.12 e 3.13 vemos que a alteração no desvio das variáveis X_1, X_3, X_4 e X_5 afeta o valor da razão $\bar{q}(A_i^*/A_0^*)$ para $i \neq 2$, fazendo com que cada razão, quando comparada ao modelo básico, fique afetada por um fator multiplicativo, aproximadamente igual ao inverso do desvio-padrão das variáveis. Construimos todas as variáveis com o mesmo desvio-padrão, igual a $\frac{1}{2}$ no modelo apresentado na tabela 3.12 e igual a 2 no modelo apresentado na tabela 3.13.

Tabela 3.12

Modelo supondo colinearidade crescente entre X_1 e X_2, sendo X_1, X_3, X_4 e $X_5 \sim N(0, \frac{1}{2}^2)$, $\epsilon \sim N(0,1)$ e $\epsilon_1 \sim N(0,1)$.

$f = 3$	$f = 3/2$	$f = 1$
$\bar{q}(A_1^*/A_0^*) = 2.1070$	$\bar{q}(A_1^*/A_0^*) = 2.6222$	$\bar{q}(A_1^*/A_0^*) = 2.8047$
$\bar{q}(A_2^*/A_0^*) = 0.3719$	$\bar{q}(A_2^*/A_0^*) = 0.6553$	$\bar{q}(A_2^*/A_0^*) = 1.0372$
$\bar{q}(A_3^*/A_0^*) = 1.9198$	$\bar{q}(A_3^*/A_0^*) = 1.8230$	$\bar{q}(A_3^*/A_0^*) = 1.7348$
$\bar{q}(A_4^*/A_0^*) = 2.2249$	$\bar{q}(A_4^*/A_0^*) = 2.1286.$	$\bar{q}(A_4^*/A_0^*) = 1.8239$
$\bar{q}(A_5^*/A_0^*) = 1.6780$	$\bar{q}(A_5^*/A_0^*) = 1.7538$	$\bar{q}(A_5^*/A_0^*) = 1.9908$

$f = 2/3$	$f = 1/3$	$f = 1/6$
$\bar{q}(A_1^*/A_0^*) = 2.1180$	$\bar{q}(A_1^*/A_0^*) = 3.3403$	$\bar{q}(A_1^*/A_0^*) = 6.3677$
$\bar{q}(A_2^*/A_0^*) = 1.2291$	$\bar{q}(A_2^*/A_0^*) = 2.9474$	$\bar{q}(A_2^*/A_0^*) = 6.8857$
$\bar{q}(A_3^*/A_0^*) = 2.0607$	$\bar{q}(A_3^*/A_0^*) = 1.8459$	$\bar{q}(A_3^*/A_0^*) = 1.7529$
$\bar{q}(A_4^*/A_0^*) = 1.5823$	$\bar{q}(A_4^*/A_0^*) = 2.0858$	$\bar{q}(A_4^*/A_0^*) = 2.4394$
$\bar{q}(A_5^*/A_0^*) = 1.7628$	$\bar{q}(A_5^*/A_0^*) = 1.7105$	$\bar{q}(A_5^*/A_0^*) = 1.9803$

$f = 1/12$	$f = 1/24$	$f = 1/48$
$\bar{q}(A_1^*/A_0^*) = 9.1705$	$\bar{q}(A_1^*/A_0^*) = 25.497$	$\bar{q}(A_1^*/A_0^*) = 52.290$
$\bar{q}(A_2^*/A_0^*) = 9.2755$	$\bar{q}(A_2^*/A_0^*) = 24.915$	$\bar{q}(A_2^*/A_0^*) = 52.483$
$\bar{q}(A_3^*/A_0^*) = 1.8998$	$\bar{q}(A_3^*/A_0^*) = 2.6667$	$\bar{q}(A_3^*/A_0^*) = 2.1606$
$\bar{q}(A_4^*/A_0^*) = 1.7686$	$\bar{q}(A_4^*/A_0^*) = 1.4198$	$\bar{q}(A_4^*/A_0^*) = 1.9568$
$\bar{q}(A_5^*/A_0^*) = 1.4914$	$\bar{q}(A_5^*/A_0^*) = 1.9283$	$\bar{q}(A_5^*/A_0^*) = 1.9425$

Neste caso temos $\hat{\alpha} = 0.4186$ e $\hat{\gamma}= 0.6432$ para $\bar{q}(A_1^*/A_0^*)$, fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 2.6218 \times (1 \times f)^{-0.6432}$$

e para $\bar{q}(A_2^*/A_0^*)$ temos $\hat{\alpha} = -0.003$ e $\hat{\gamma}= 0.9983$, fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 0.9932 \times (1 \times f)^{-0.9983}$$

A análise da tabela 3.12 nos mostra que os valores de $\bar{q}(A_i^*/A_0^*)$ para $i \neq 2$ ficam alterados com a modificação feita no desvio-padrão de X_i. Entretanto, a medida que o grau de colinearidade aumenta a influência do desvio-padrão de X_i no valor de $\bar{q}(A_2^*/A_0^*)$ vai ficando menos visível, enquanto que, para as variáveis que não estão envolvidas em colinearidade o valor do patamar, que no modelo básico era de 1, fica próximo de 2, nos levando a sugerir que este patamar seria dado pelo inverso do desvio-padrão comum de X_i. Podemos observar também que a colinearidade, construída por $X_2 = X_1 + f\epsilon_1$, começa ficar evidente quando $f < \sigma_{X_1}$.

Tabela 3.13

Modelo supondo colinearidade crescente entre X_1 e X_2,
sendo X_1, X_3, X_4 e $X_5 \sim N(0, 2^2)$, $\epsilon \sim N(0,1)$ e $\epsilon_1 \sim N(0,1)$.

$f = 3$	$f = 3/2$	$f = 1$
$\bar{q}(A_1^*/A_0^*) = 0.4637$	$\bar{q}(A_1^*/A_0^*) = 0.8693$	$\bar{q}(A_1^*/A_0^*) = 1.1543$
$\bar{q}(A_2^*/A_0^*) = 0.2466$	$\bar{q}(A_2^*/A_0^*) = 0.6884$	$\bar{q}(A_2^*/A_0^*) = 1.0558$
$\bar{q}(A_3^*/A_0^*) = 0.4232$	$\bar{q}(A_3^*/A_0^*) = 0.5502$	$\bar{q}(A_3^*/A_0^*) = 0.5222$
$\bar{q}(A_4^*/A_0^*) = 0.6445$	$\bar{q}(A_4^*/A_0^*) = 0.3675$	$\bar{q}(A_4^*/A_0^*) = 0.5981$
$\bar{q}(A_5^*/A_0^*) = 0.3871$	$\bar{q}(A_5^*/A_0^*) = 0.4015$	$\bar{q}(A_5^*/A_0^*) = 0.5681$

$f = 2/3$	$f = 1/3$	$f = 1/6$
$\bar{q}(A_1^*/A_0^*) = 1.4815$	$\bar{q}(A_1^*/A_0^*) = 2.8166$	$\bar{q}(A_1^*/A_0^*) = 5.9511$
$\bar{q}(A_2^*/A_0^*) = 1.2709$	$\bar{q}(A_2^*/A_0^*) = 2.6783$	$\bar{q}(A_2^*/A_0^*) = 5.9944$
$\bar{q}(A_3^*/A_0^*) = 0.5121$	$\bar{q}(A_3^*/A_0^*) = 0.5617$	$\bar{q}(A_3^*/A_0^*) = 0.6222$
$\bar{q}(A_4^*/A_0^*) = 0.5021$	$\bar{q}(A_4^*/A_0^*) = 0.4681$	$\bar{q}(A_4^*/A_0^*) = 0.5737$
$\bar{q}(A_5^*/A_0^*) = 0.3933$	$\bar{q}(A_5^*/A_0^*) = 0.5486$	$\bar{q}(A_5^*/A_0^*) = 0.5640$

$f = 1/12$	$f = 1/24$	$f = 1/48$
$\bar{q}(A_1^*/A_0^*) = 10.531$	$\bar{q}(A_1^*/A_0^*) = 20.705$	$\bar{q}(A_1^*/A_0^*) = 36.010$
$\bar{q}(A_2^*/A_0^*) = 10.597$	$\bar{q}(A_2^*/A_0^*) = 20.918$	$\bar{q}(A_2^*/A_0^*) = 36.060$
$\bar{q}(A_3^*/A_0^*) = 0.4425$	$\bar{q}(A_3^*/A_0^*) = 0.4363$	$\bar{q}(A_3^*/A_0^*) = 0.3810$
$\bar{q}(A_4^*/A_0^*) = 0.6189$	$\bar{q}(A_4^*/A_0^*) = 0.5633$	$\bar{q}(A_4^*/A_0^*) = 0.4501$
$\bar{q}(A_5^*/A_0^*) = 0.4782$	$\bar{q}(A_5^*/A_0^*) = 0.6105$	$\bar{q}(A_5^*/A_0^*) = 0.4893$

Neste caso temos $\hat{\alpha} = 0.0656$ e $\hat{\gamma}= 0.8877$ para $\bar{q}(A_1^*/A_0^*)$, fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 1.1631 \times (1 \times f)^{-0.8877}$$

e para $\bar{q}(A_2^*/A_0^*)$ temos $\hat{\alpha} = -0.0404$ e $\hat{\gamma}= 0.9819$, fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 0.9111 \times (1 \times f)^{-0.9819}$$

A análise da tabela 3.13 nos leva as mesmas observações já feitas no modelo anterior, tabela 3.12, sendo que neste caso o patamar das variáveis não envolvidas em colinearidade fica próximo de $\frac{1}{2}$, reafirmando a nossa suspeita de que esse patamar é dado pelo inverso do desvio-padrão comum dos X_i. É claro que, até esse momento a única modificação feita foi no desvio de X_i, precisamos ainda explorar outras situações.

Com as modificações feitas no desvio-padrão de ϵ_1 observamos também, algumas modificações nos valores de 10^{α} e γ nas equações de regressão. Analisando a tabela 3.14 podemos observar que as mesmas alterações ocorridas nos valores de 10^{α} e γ nos modelos avaliados no item 3.4.2.1, ocorrem novamente nesse modelos. Os valores de 10^{α} e γ sofrem as maiores alterações para a razão $\bar{q}(A_1^*/A_0^*)$ que está relacionada a variável X_1, que como dissemos anteriormente carrega o peso da variável X_2. Além disso, o menor valor de γ está associado a situação onde $\sigma_{\epsilon_1} > \sigma_{X_i}$, como já foi visto no modelo apresentado no item 3.4.2.1. Entretanto, se analisarmos a tabela 3.15, vemos que os valores de 10^{α} e γ se mantem próximos de um.

Tabela 3.14

$\bar{q}(A_1^*/A_0^*)$

σ_{X_i}	σ_{ϵ_1}	10^{α}	γ
$\frac{1}{2}$	1	2.6218	-0.6432
1	1	1.5517	-0.8569
2	1	1.1631	-0.8877

e

Tabela 3.15

$\bar{q}(A_2^*/A_0^*)$

σ_{X_i}	σ_{ϵ_1}	10^{α}	γ
$\frac{1}{2}$	1	0.9932	-0.9983
1	1	0.9505	-1.0343
2	1	0.9111	-0.9819

Nos modelos analisados até agrora vimos que tanto a alteração feita no desvio-padrão das variáveis X_1, X_3, X_4 e X_5, como a feita no desvio-padrão de ϵ_1, implicam em uma modificação no valor da razão $\bar{q}$. Vamos agora combinar esses dois fatores: desvios-padrão diferentes de 1 para as variáveis e para ϵ_1.

3.4.2.3-Alterações no Desvio das Variáveis X_1, X_3, X_4 e X_5

mantendo $\epsilon_1 \sim N\left(0, \frac{1}{2}^2\right)$

Analisando as tabelas 3.16 à 3.19 vemos que a combinação dos dois fatores, também é manifestada no valor da razão $\bar{q}$. Sendo igual a que foi apresentada nos modelos anteriores onde cada alteração foi feita separadamente, a única diferença é que para cada combinação de modificações o valor da razão $\bar{q}$ se altera combinando as duas modificações. Por exemplo, na tabela 3.15 em que o desvio-padrão de ϵ_1 é $\frac{1}{2}$, o valor de $\bar{q}(A_2^*/A_0^*)$ dobra

em relação ao modelo básico, mas como o desvio-padrão das variáveis também é $\frac{1}{2}$ o valor de $\bar{q}(A_i^*/A_0^*)$ para $i = 1, 3, 4$ e 5, dobra, isto é, no modelo básico ele é aproximadamente igual a 1 e agora passa a ser aproximadamente igual a 2, vemos novamente a alteração do patamar das variáveis que não estão envolvidas em colinearidade.

Tabela 3.16

Modelo supondo colinearidade crescente entre as variáveis X_1 e X_2, sendo X_1, X_3, X_4 e $X_5 \sim N(0, \frac{1}{2}^2)$, $\epsilon \sim N(0,1)$ e $\epsilon_1 \sim N(0, \frac{1}{2}^2)$.

$f = 3$	$f = 3/2$	$f = 1$
$\bar{q}(A_1^*/A_0^*) = 2.0703$	$\bar{q}(A_1^*/A_0^*) = 2.1140$	$\bar{q}(A_1^*/A_0^*) = 2.3766$
$\bar{q}(A_2^*/A_0^*) = 0.7576$	$\bar{q}(A_2^*/A_0^*) = 1.3097$	$\bar{q}(A_2^*/A_0^*) = 1.4503$
$\bar{q}(A_3^*/A_0^*) = 1.7958$	$\bar{q}(A_3^*/A_0^*) = 1.8955$	$\bar{q}(A_3^*/A_0^*) = 2.0657$
$\bar{q}(A_4^*/A_0^*) = 2.1102$	$\bar{q}(A_4^*/A_0^*) = 1.8006$	$\bar{q}(A_4^*/A_0^*) = 2.0485$
$\bar{q}(A_5^*/A_0^*) = 1.6956$	$\bar{q}(A_5^*/A_0^*) = 1.7019$	$\bar{q}(A_5^*/A_0^*) = 2.7834$

$f = 2/3$	$f = 1/3$	$f = 1/6$
$\bar{q}(A_1^*/A_0^*) = 4.7417$	$\bar{q}(A_1^*/A_0^*) = 5.7541$	$\bar{q}(A_1^*/A_0^*) = 10.207$
$\bar{q}(A_2^*/A_0^*) = 4.0663$	$\bar{q}(A_2^*/A_0^*) = 5.5447$	$\bar{q}(A_2^*/A_0^*) = 9.9989$
$\bar{q}(A_3^*/A_0^*) = 2.3167$	$\bar{q}(A_3^*/A_0^*) = 1.9171$	$\bar{q}(A_3^*/A_0^*) = 1.6680$
$\bar{q}(A_4^*/A_0^*) = 2.0342$	$\bar{q}(A_4^*/A_0^*) = 2.2222$	$\bar{q}(A_4^*/A_0^*) = 1.9392$
$\bar{q}(A_5^*/A_0^*) = 2.4286$	$\bar{q}(A_5^*/A_0^*) = 1.7364$	$\bar{q}(A_5^*/A_0^*) = 1.7782$

$f = 1/12$	$f = 1/24$	$f = 1/48$
$\bar{q}(A_1^*/A_0^*) = 25.441$	$\bar{q}(A_1^*/A_0^*) = 48.765$	$\bar{q}(A_1^*/A_0^*) = 96.372$
$\bar{q}(A_2^*/A_0^*) = 25.322$	$\bar{q}(A_2^*/A_0^*) = 48.272$	$\bar{q}(A_2^*/A_0^*) = 97.091$
$\bar{q}(A_3^*/A_0^*) = 2.4077$	$\bar{q}(A_3^*/A_0^*) = 2.2623$	$\bar{q}(A_3^*/A_0^*) = 2.1398$
$\bar{q}(A_4^*/A_0^*) = 2.6648$	$\bar{q}(A_4^*/A_0^*) = 1.7945$	$\bar{q}(A_4^*/A_0^*) = 3.1303$
$\bar{q}(A_5^*/A_0^*) = 1.8095$	$\bar{q}(A_5^*/A_0^*) = 2.1354$	$\bar{q}(A_5^*/A_0^*) = 2.4589$

Neste caso temos $\hat{\alpha} = 0.2494$ e $\hat{\gamma}= 0.8252$ para $\bar{q}(A_1^*/A_0^*)$, fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 1.7758 \times (\frac{1}{2} \times f)^{-0.8252}$$

e para $\bar{q}(A_2^*/A_0^*)$ temos $\hat{\alpha} = -0.0008$ e $\hat{\gamma}= 0.9983$, fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 0.9980 \times (\frac{1}{2} \times f)^{-0.9983}$$

Novamente notamos a alteração dos valores de 10^{α} e γ na equação de regressão ajustada para os valores de $\bar{q}(A_1^*/A_0^*)$. Vamos comparar esses valores com os resultados observados no modelo da tabela 3.17.

Tabela 3.17

Modelo supondo colinearidade crescente entre X_1 e X_2,
sendo X_1, X_3, X_4 e $X_5 \sim N(0, 2^2)$, $\epsilon \sim N(0,1)$ e $\epsilon_1 \sim N(0, \frac{1}{2}^2)$.

$f = 3$	$f = 3/2$	$f = 1$
$\bar{q}(A_1^*/A_0^*) = 1.1131$	$\bar{q}(A_1^*/A_0^*) = 1.3314$	$\bar{q}(A_1^*/A_0^*) = 2.2936$
$\bar{q}(A_2^*/A_0^*) = 0.9223$	$\bar{q}(A_2^*/A_0^*) = 1.3093$	$\bar{q}(A_2^*/A_0^*) = 2.3327$
$\bar{q}(A_3^*/A_0^*) = 0.5100$	$\bar{q}(A_3^*/A_0^*) = 0.5168$	$\bar{q}(A_3^*/A_0^*) = 0.4943$
$\bar{q}(A_4^*/A_0^*) = 0.7632$	$\bar{q}(A_4^*/A_0^*) = 0.6534$	$\bar{q}(A_4^*/A_0^*) = 0.4619$
$\bar{q}(A_5^*/A_0^*) = 0.5483$	$\bar{q}(A_5^*/A_0^*) = 0.4960$	$\bar{q}(A_5^*/A_0^*) = 0.4752$

A tabela continua na página seguinte

$f = 2/3$	$f = 1/3$	$f = 1/6$
$\bar{q}(A_1^*/A_0^*) = 2.7381$	$\bar{q}(A_1^*/A_0^*) = 6.8921$	$\bar{q}(A_1^*/A_0^*) = 8.4548$
$\bar{q}(A_2^*/A_0^*) = 2.7005$	$\bar{q}(A_2^*/A_0^*) = 6.9970$	$\bar{q}(A_2^*/A_0^*) = 8.5615$
$\bar{q}(A_3^*/A_0^*) = 0.6172$	$\bar{q}(A_3^*/A_0^*) = 0.5465$	$\bar{q}(A_3^*/A_0^*) = 0.3521$
$\bar{q}(A_4^*/A_0^*) = 0.6749$	$\bar{q}(A_4^*/A_0^*) = 0.8278$	$\bar{q}(A_4^*/A_0^*) = 0.3828$
$\bar{q}(A_5^*/A_0^*) = 0.6215$	$\bar{q}(A_5^*/A_0^*) = 0.5263$	$\bar{q}(A_5^*/A_0^*) = 0.5877$

$f = 1/12$	$f = 1/24$	$f = 1/48$
$\bar{q}(A_1^*/A_0^*) = 24.002$	$\bar{q}(A_1^*/A_0^*) = 48.638$	$\bar{q}(A_1^*/A_0^*) = 110.76$
$\bar{q}(A_2^*/A_0^*) = 23.945$	$\bar{q}(A_2^*/A_0^*) = 49.042$	$\bar{q}(A_2^*/A_0^*) = 110.92$
$\bar{q}(A_3^*/A_0^*) = 0.6332$	$\bar{q}(A_3^*/A_0^*) = 0.5735$	$\bar{q}(A_3^*/A_0^*) = 0.4906$
$\bar{q}(A_4^*/A_0^*) = 0.5213$	$\bar{q}(A_4^*/A_0^*) = 0.5596$	$\bar{q}(A_4^*/A_0^*) = 0.5670$
$\bar{q}(A_5^*/A_0^*) = 0.4665$	$\bar{q}(A_5^*/A_0^*) = 0.5518$	$\bar{q}(A_5^*/A_0^*) = 0.5764$

Neste caso temos $\hat{\alpha} = 0.0652$ e $\hat{\gamma}= 0.9517$, fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 1.1620 \times (\frac{1}{2} \times f)^{-0.9517}$$

e para $\bar{q}(A_2^*/A_0^*)$ temos $\hat{\alpha} = -0.0393$ e $\hat{\gamma}= 0.9727$, fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 1.0948 \times (\frac{1}{2} \times f)^{-0.9727}$$

Tabela 3.18

$\bar{q}(A_1^*/A_0^*)$

σ_{X_i}	σ_{ϵ_1}	10^{α}	γ
$\frac{1}{2}$	$\frac{1}{2}$	1.7758	-0.8258
2	$\frac{1}{2}$	1.1620	-0.9517

e

Tabela 3.19

$\bar{q}(A_2^*/A_0^*)$

σ_{X_i}	σ_{ϵ_1}	10^{α}	γ
$\frac{1}{2}$	$\frac{1}{2}$	0.9980	-0.9983
2	$\frac{1}{2}$	1.0948	-0.9727

Analisando as tabelas 3.18 e 3.19 vemos que as alterações nos valores de 10^{α} e γ, é observada para $\bar{q}(A_1^*/A_0^*)$, sendo que neste caso, as alterações são menos significativas do que as observadas anteriormente nos modelos apresentados nos itens 3.4.2.1 à 3.4.2.3. Nos dois modelos σ_{ϵ_1} é menor ou igual a σ_{X_i}, não influenciando tanto os valores de 10^{α} e γ, entretantro uma leve alteração ainda pode ser observada ao compararmos os valores de 10^{α} e γ nas tabelas 3.18 e 3.19, podendo ser atribuida, mais uma vez, ao peso que a variável X_1 carrega na construcão da variável X_2.

3.4.2.4-Alterações no Desvio das Variáveis X_1, X_3, X_4 e X_5 mantendo $\epsilon_1 \sim N(0, 2^2)$

Tabela 3.20

Modelo supondo colinearidade crescente entre X_1 e X_2,
sendo X_1, X_3, X_4 e $X_5 \sim N(0, \frac{1}{2}^2)$, $\epsilon \sim N(0,1)$ e $\epsilon_1 \sim N(0, 2^2)$.

$f = 3$
$\bar{q}(A_1^*/A_0^*) = 1.8131$
$\bar{q}(A_2^*/A_0^*) = 0.1327$
$\bar{q}(A_3^*/A_0^*) = 1.9436$
$\bar{q}(A_4^*/A_0^*) = 2.3542$
$\bar{q}(A_5^*/A_0^*) = 1.6458$

$f = 3/2$
$\bar{q}(A_1^*/A_0^*) = 2.1484$
$\bar{q}(A_2^*/A_0^*) = 0.3271$
$\bar{q}(A_3^*/A_0^*) = 1.8559$
$\bar{q}(A_4^*/A_0^*) = 2.2931$
$\bar{q}(A_5^*/A_0^*) = 1.5908$

$f = 1$
$\bar{q}(A_1^*/A_0^*) = 2.3175$
$\bar{q}(A_2^*/A_0^*) = 0.5893$
$\bar{q}(A_3^*/A_0^*) = 1.9737$
$\bar{q}(A_4^*/A_0^*) = 2.3596$
$\bar{q}(A_5^*/A_0^*) = 2.1153$

$f = 2/3$
$\bar{q}(A_1^*/A_0^*) = 3.1553$
$\bar{q}(A_2^*/A_0^*) = 0.8161$
$\bar{q}(A_3^*/A_0^*) = 2.3034$
$\bar{q}(A_4^*/A_0^*) = 1.8493$
$\bar{q}(A_5^*/A_0^*) = 2.4790$

$f = 1/3$
$\bar{q}(A_1^*/A_0^*) = 2.1088$
$\bar{q}(A_2^*/A_0^*) = 1.3087$
$\bar{q}(A_3^*/A_0^*) = 1.6148$
$\bar{q}(A_4^*/A_0^*) = 2.0372$
$\bar{q}(A_5^*/A_0^*) = 2.3677$

$f = 1/6$
$\bar{q}(A_1^*/A_0^*) = 2.9779$
$\bar{q}(A_2^*/A_0^*) = 2.4990$
$\bar{q}(A_3^*/A_0^*) = 1.7763$
$\bar{q}(A_4^*/A_0^*) = 1.8377$
$\bar{q}(A_5^*/A_0^*) = 1.6504$

$f = 1/12$
$\bar{q}(A_1^*/A_0^*) = 6.5068$
$\bar{q}(A_2^*/A_0^*) = 6.6046$
$\bar{q}(A_3^*/A_0^*) = 2.4244$
$\bar{q}(A_4^*/A_0^*) = 2.2346$
$\bar{q}(A_5^*/A_0^*) = 2.4765$

$f = 1/24$
$\bar{q}(A_1^*/A_0^*) = 10.631$
$\bar{q}(A_2^*/A_0^*) = 10.456$
$\bar{q}(A_3^*/A_0^*) = 2.4246$
$\bar{q}(A_4^*/A_0^*) = 2.2848$
$\bar{q}(A_5^*/A_0^*) = 1.7985$

$f = 1/48$
$\bar{q}(A_1^*/A_0^*) = 21.538$
$\bar{q}(A_2^*/A_0^*) = 20.842$
$\bar{q}(A_3^*/A_0^*) = 1.3125$
$\bar{q}(A_4^*/A_0^*) = 1.7617$
$\bar{q}(A_5^*/A_0^*) = 2.0297$

A tabela 3.20 nos mostra mais uma vez a influência que sofrem os valores de $\bar{q}(A_i^*/A_0^*)$ quando alterações são feitas no desvio de ϵ_1 e das variáveis.

Neste caso temos $\hat{\alpha} = 0.4895$ e $\hat{\gamma}= 0.4619$ para $\bar{q}(A_1^*/A_0^*)$, fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 3.0867 \times (2 \times f)^{-0.4619}$$

e para $\bar{q}(A_2^*/A_0^*)$ temos $\hat{\alpha} = -0.0209$ e $\hat{\gamma}= 0.9869$, fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 0.9530 \times (2 \times f)^{-0.9869}$$

Ao compararmos as equações de regressão ajustadas para $\bar{q}(A_i^*/A_0^*)$ $i = 1$ e 2, vemos que os valores de 10^{α} e γ estão bem diferentes de 1 em $\bar{q}(A_1^*/A_0^*)$, e não sofrem quase nenhuma modificação para a equação ajustada para os valores de $\bar{q}(A_2^*/A_0^*)$.

Tabela 3.21

Modelo supondo colinearidade crescente entre X_1 e X_2,
sendo X_1, X_3, X_4 e $X_5 \sim N(0, 2^2)$, $\epsilon \sim N(0,1)$ e $\epsilon_1 \sim N(0, 2^2)$.

$f = 3$	$f = 3/2$	$f = 1$
$\bar{q}(A_1^*/A_0^*) = 0.5676$	$\bar{q}(A_1^*/A_0^*) = 0.7808$	$\bar{q}(A_1^*/A_0^*) = 0.8722$
$\bar{q}(A_2^*/A_0^*) = 0.2008$	$\bar{q}(A_2^*/A_0^*) = 0.4650$	$\bar{q}(A_2^*/A_0^*) = 0.4089$
$\bar{q}(A_3^*/A_0^*) = 0.6423$	$\bar{q}(A_3^*/A_0^*) = 0.4864$	$\bar{q}(A_3^*/A_0^*) = 0.4581$
$\bar{q}(A_4^*/A_0^*) = 0.5704$	$\bar{q}(A_4^*/A_0^*) = 0.5295$	$\bar{q}(A_4^*/A_0^*) = 0.4173$
$\bar{q}(A_5^*/A_0^*) = 0.6544$	$\bar{q}(A_5^*/A_0^*) = 0.6875$	$\bar{q}(A_5^*/A_0^*) = 0.4057$

A tabela continua na página seguinte

$f = 2/3$	$f = 1/3$	$f = 1/6$
$\bar{q}(A_1^*/A_0^*) = 0.8784$	$\bar{q}(A_1^*/A_0^*) = 1.9569$	$\bar{q}(A_1^*/A_0^*) = 2.8159$
$\bar{q}(A_2^*/A_0^*) = 0.7773$	$\bar{q}(A_2^*/A_0^*) = 1.9568$	$\bar{q}(A_2^*/A_0^*) = 2.8190$
$\bar{q}(A_3^*/A_0^*) = 0.5256$	$\bar{q}(A_3^*/A_0^*) = 0.4822$	$\bar{q}(A_3^*/A_0^*) = 0.5335$
$\bar{q}(A_4^*/A_0^*) = 0.5959$	$\bar{q}(A_4^*/A_0^*) = 0.6626$	$\bar{q}(A_4^*/A_0^*) = 0.7057$
$\bar{q}(A_5^*/A_0^*) = 0.4261$	$\bar{q}(A_5^*/A_0^*) = 0.5139$	$\bar{q}(A_5^*/A_0^*) = 0.5039$

$f = 1/12$	$f = 1/24$	$f = 1/48$
$\bar{q}(A_1^*/A_0^*) = 5.7454$	$\bar{q}(A_1^*/A_0^*) = 13.109$	$\bar{q}(A_1^*/A_0^*) = 19.161$
$\bar{q}(A_2^*/A_0^*) = 5.8451$	$\bar{q}(A_2^*/A_0^*) = 12.917$	$\bar{q}(A_2^*/A_0^*) = 19.157$
$\bar{q}(A_3^*/A_0^*) = 0.3601$	$\bar{q}(A_3^*/A_0^*) = 0.4586$	$\bar{q}(A_3^*/A_0^*) = 0.4289$
$\bar{q}(A_4^*/A_0^*) = 0.3799$	$\bar{q}(A_4^*/A_0^*) = 0.8342$	$\bar{q}(A_4^*/A_0^*) = 0.6397$
$\bar{q}(A_5^*/A_0^*) = 0.3988$	$\bar{q}(A_5^*/A_0^*) = 0.6330$	$\bar{q}(A_5^*/A_0^*) = 0.5053$

Neste caso temos $\hat{\alpha} = 0.1966$ e $\hat{\gamma}= 0.7523$ para $\bar{q}(A_1^*/A_0^*)$, fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 1.5725 \times (2 \times f)^{-0.7523}$$

e para $\bar{q}(A_2^*/A_0^*)$ temos $\hat{\alpha} = 0.0325$ e $\hat{\gamma}= 0.9423$, fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 1.0777 \times (2 \times f)^{-0.9423}$$

Tabela 3.22

$\bar{q}(A_1^*/A_0^*)$

σ_{X_i}	σ_{ϵ_1}	10^{α}	γ
$\frac{1}{2}$	2	3.0867	-0.4619
2	2	1.5725	-0.7523

e

Tabela 3.23

$\bar{q}(A_2^*/A_0^*)$

σ_{X_i}	σ_{ϵ_1}	10^{α}	γ
$\frac{1}{2}$	2	0.9530	-0.9869
2	2	1.0777	-0.9423

As tabelas 3.22 e 3.23 mostram mais uma vez as alterações sofridas nos valores de 10^{α} e γ para $\bar{q}(A_1^*/A_0^*)$, sendo que quanto maior o valor de σ_{ϵ_1} em relação ao valor de σ_{X_i}, mais distantes de 1 ficam os valores de 10^{α} e γ.

Com todas essas modificações feitas nos modelos, vimos que independente da alteração feita, nos desvios-padrão das variáveis X_i e de ϵ_1, a razão $\bar{q}$ continua a mostrar quais as variáveis estão envolvidas em colinearidade. Observamos que o valor das razões $\bar{q}$ para as variáveis envolvidas na colinearidade, está relacionado ao fator de colinearidade entre estas variáveis e ao desvio de ϵ_1, e o patamar apresentado pelas variáveis não envolvidas em colinearidade está relacionado com o desvio-padrão das variáveis X_1, X_3, X_4 e X_5. Podemos observar também que a colinearidade, construida por $X_2 = X_1 + f\epsilon_1$, começa a ficar evidente quando $f < \sigma_{X_1}$.

Até o momento todas as variáveis envolvidas nos modelos apresentavam o mesmo desvio-padrão, que como pudemos observar influência o valor da razão $\bar{q}$. Decidimos então investigar o comportamento de $\bar{q}$ quando os desvios-padrão são diferentes para

cada uma das variáveis, e além disso padronizamos as variáveis com o objetivo de não mascarar o efeito da colinearidade nas variáveis, pois vimos no modelo apresentado na tabela 3.11, que o valor do desvio-padrão das variáveis X_i influência o valor da razão $\bar{q}$.

Modelos onde as variáveis X_1, X_3, X_4 e X_5 tem desvios diferentes

Construimos a princípio um modelo de referência onde não existe colinearidade, mas as variáveis tem desvios-padrão diferentes e para o cálculo da razão $\bar{q}$ as variáveis foram padronizadas.

Tabela 3.24

Modelo sem colinearidade,
supondo: $X_1 \sim N(0, 5.5^2)$,$X_2 \sim N(0, 2.7^2)$,
$X_3 \sim N(0, 1.2^2)$,$X_4 \sim N(0, 0.7^2)$ e $X_5 \sim N(0, 1.8^2)$.

referência
$\bar{q}(A_1^*/A_0^*) = 1.0036$
$\bar{q}(A_2^*/A_0^*) = 1.0061$
$\bar{q}(A_3^*/A_0^*) = 1.0100$
$\bar{q}(A_4^*/A_0^*) = 1.0115$
$\bar{q}(A_5^*/A_0^*) = 0.8649$

Ao analisarmos as tabelas 3.25 à 3.27, vemos que já para o primeiro fator de colinearidade, $f = 3$, os valores de $\bar{q}(A_i^*/A_0^*)$ $i = 1$ e 2, se destacam dos demais, mostrando que X_1 e X_2 são as variáveis envolvidas em colinearidade. Vemos também que o desvio-padrão de ϵ_1 influência os valores de $\bar{q}(A_i^*/A_0^*)$ $i = 1$ e 2, isto é, se compararmos estes valores para os três modelos, tomando como referência o modelo onde o desvio-padrão de ϵ_1 é 1 (tabela 3.26), temos que para o modelo onde o desvio-padrão de ϵ_1 é $\frac{1}{2}$, os valores de $\bar{q}(A_i^*/A_0^*)$ $i = 1$ e 2, ficam multiplicados por dois, e para o modelo onde o desvio de ϵ_1 é 2 os valores ficam divididos por 2.

Tabela 3.25

Modelo supondo colinearidade crescente entre X_1 e X_2,
sendo $X_1 \sim N(0, 5.5^2)$, $X_3 \sim N(0, 1.2^2)$, $X_4 \sim N(0, 0.7^2)$, $X_5 \sim N(0, 1.8^2)$,
$\epsilon \sim N(0,1)$ e $\epsilon_1 \sim N(0, \frac{1}{2}^2)$.

$f = 3$
$\bar{q}(A_1^*/A_0^*) = 3.8348$
$\bar{q}(A_2^*/A_0^*) = 3.8673$
$\bar{q}(A_3^*/A_0^*) = 0.9735$
$\bar{q}(A_4^*/A_0^*) = 1.2085$
$\bar{q}(A_5^*/A_0^*) = 1.0168$

$f = 3/2$
$\bar{q}(A_1^*/A_0^*) = 8.6847$
$\bar{q}(A_2^*/A_0^*) = 8.4871$
$\bar{q}(A_3^*/A_0^*) = 1.1920$
$\bar{q}(A_4^*/A_0^*) = 0.9570$
$\bar{q}(A_5^*/A_0^*) = 0.9546$

$f = 1$
$\bar{q}(A_1^*/A_0^*) = 18.083$
$\bar{q}(A_2^*/A_0^*) = 18.174$
$\bar{q}(A_3^*/A_0^*) = 1.0840$
$\bar{q}(A_4^*/A_0^*) = 1.1071$
$\bar{q}(A_5^*/A_0^*) = 1.1002$

$f = 2/3$
$\bar{q}(A_1^*/A_0^*) = 22.999$
$\bar{q}(A_2^*/A_0^*) = 22.715$
$\bar{q}(A_3^*/A_0^*) = 0.9024$
$\bar{q}(A_4^*/A_0^*) = 1.1236$
$\bar{q}(A_5^*/A_0^*) = 0.8931$

$f = 1/3$
$\bar{q}(A_1^*/A_0^*) = 33.306$
$\bar{q}(A_2^*/A_0^*) = 32.887$
$\bar{q}(A_3^*/A_0^*) = 0.9447$
$\bar{q}(A_4^*/A_0^*) = 0.9774$
$\bar{q}(A_5^*/A_0^*) = 0.9133$

$f = 1/6$
$\bar{q}(A_1^*/A_0^*) = 66.679$
$\bar{q}(A_2^*/A_0^*) = 67.271$
$\bar{q}(A_3^*/A_0^*) = 1.1344$
$\bar{q}(A_4^*/A_0^*) = 1.1568$
$\bar{q}(A_5^*/A_0^*) = 1.0644$

$f = 1/12$
$\bar{q}(A_1^*/A_0^*) = 120.95$
$\bar{q}(A_2^*/A_0^*) = 121.27$
$\bar{q}(A_3^*/A_0^*) = 0.9977$
$\bar{q}(A_4^*/A_0^*) = 1.0030$
$\bar{q}(A_5^*/A_0^*) = 0.9703$

$f = 1/24$
$\bar{q}(A_1^*/A_0^*) = 264.34$
$\bar{q}(A_2^*/A_0^*) = 263.47$
$\bar{q}(A_3^*/A_0^*) = 1.3882$
$\bar{q}(A_4^*/A_0^*) = 1.1096$
$\bar{q}(A_5^*/A_0^*) = 1.4679$

$f = 1/48$
$\bar{q}(A_1^*/A_0^*) = 547.56$
$\bar{q}(A_2^*/A_0^*) = 514.65$
$\bar{q}(A_3^*/A_0^*) = 1.1282$
$\bar{q}(A_4^*/A_0^*) = 0.9627$
$\bar{q}(A_5^*/A_0^*) = 0.9544$

Neste caso temos $\hat{\alpha} = 0.8373$ e $\hat{\gamma}= 0.9433$ para $\bar{q}(A_1^*/A_0^*)$, fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 6.8754 \times (\frac{1}{2} \times f)^{-0.9433}$$

e para $\bar{q}(A_2^*/A_0^*)$ temos $\hat{\alpha} = 0.8380$ e $\hat{\gamma}= 0.9374$, fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 6.8869 \times (\frac{1}{2} \times f)^{-0.9374}$$

Tabela 3.26

Modelo supondo colinearidade crescente entre X_1 e X_2,
sendo $X_1 \sim N(0, 5.5^2)$, $X_3 \sim N(0, 1.2^2)$, $X_4 \sim N(0, 0.7^2)$, $X_5 \sim N(0, 1.8^2)$,
$\epsilon \sim N(0,1)$ e $\epsilon_1 \sim N(0,1)$.

$f = 3$	$f = 3/2$	$f = 1$
$\bar{q}(A_1^*/A_0^*) = 2.1367$	$\bar{q}(A_1^*/A_0^*) = 4.1319$	$\bar{q}(A_1^*/A_0^*) = 9.1683$
$\bar{q}(A_2^*/A_0^*) = 2.1448$	$\bar{q}(A_2^*/A_0^*) = 4.0728$	$\bar{q}(A_2^*/A_0^*) = 9.1734$
$\bar{q}(A_3^*/A_0^*) = 0.9735$	$\bar{q}(A_3^*/A_0^*) = 1.1650$	$\bar{q}(A_3^*/A_0^*) = 1.0839$
$\bar{q}(A_4^*/A_0^*) = 1.2376$	$\bar{q}(A_4^*/A_0^*) = 0.9967$	$\bar{q}(A_4^*/A_0^*) = 1.1071$
$\bar{q}(A_5^*/A_0^*) = 1.0168$	$\bar{q}(A_5^*/A_0^*) = 0.9544$	$\bar{q}(A_5^*/A_0^*) = 1.1002$

A tabela continua na página seguinte

$f = 2/3$	$f = 1/3$	$f = 1/6$
$\bar{q}(A_1^*/A_0^*) = 11.692$	$\bar{q}(A_1^*/A_0^*) = 16.894$	$\bar{q}(A_1^*/A_0^*) = 33.348$
$\bar{q}(A_2^*/A_0^*) = 11.451$	$\bar{q}(A_2^*/A_0^*) = 16.499$	$\bar{q}(A_2^*/A_0^*) = 33.534$
$\bar{q}(A_3^*/A_0^*) = 0.9024$	$\bar{q}(A_3^*/A_0^*) = 0.9447$	$\bar{q}(A_3^*/A_0^*) = 1.1344$
$\bar{q}(A_4^*/A_0^*) = 1.1236$	$\bar{q}(A_4^*/A_0^*) = 0.9774$	$\bar{q}(A_4^*/A_0^*) = 1.1568$
$\bar{q}(A_5^*/A_0^*) = 0.8931$	$\bar{q}(A_5^*/A_0^*) = 0.9133$	$\bar{q}(A_5^*/A_0^*) = 1.0644$

$f = 1/12$	$f = 1/24$	$f = 1/48$
$\bar{q}(A_1^*/A_0^*) = 60.472$	$\bar{q}(A_1^*/A_0^*) = 132.22$	$\bar{q}(A_1^*/A_0^*) = 257.24$
$\bar{q}(A_2^*/A_0^*) = 60.590$	$\bar{q}(A_2^*/A_0^*) = 131.74$	$\bar{q}(A_2^*/A_0^*) = 257.38$
$\bar{q}(A_3^*/A_0^*) = 0.9977$	$\bar{q}(A_3^*/A_0^*) = 1.3882$	$\bar{q}(A_3^*/A_0^*) = 1.1282$
$\bar{q}(A_4^*/A_0^*) = 1.0030$	$\bar{q}(A_4^*/A_0^*) = 1.1096$	$\bar{q}(A_4^*/A_0^*) = 0.9627$
$\bar{q}(A_5^*/A_0^*) = 0.9703$	$\bar{q}(A_5^*/A_0^*) = 1.4679$	$\bar{q}(A_5^*/A_0^*) = 0.9544$

Neste caso $\hat{\alpha} = 0.8310$ e $\hat{\gamma}= 0.9273$ para $\bar{q}(A_1^*/A_0^*)$, fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 6.7764 \times (1 \times f)^{-0.9273}$$

e para $\bar{q}(A_2^*/A_0^*)$ temos $\hat{\alpha} = 0.8277$ e $\hat{\gamma}= 0.9288$, fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 6.7258 \times (1 \times f)^{-0.9288}$$

Tabela 3.27

Modelo supondo colinearidade crescente entre X_1 e X_2,
sendo $X_1 \sim N(0, 5.5^2)$, $X_3 \sim N(0, 1.2^2)$, $X_4 \sim N(0, 0.7^2)$, $X_5 \sim N(0, 1.8^2)$,
$\epsilon \sim N(0,1)$ e $\epsilon_1 \sim N(0, 2^2)$.

$f = 3$
$\bar{q}(A_1^*/A_0^*) = 1.4005$
$\bar{q}(A_2^*/A_0^*) = 1.3893$
$\bar{q}(A_3^*/A_0^*) = 0.9735$
$\bar{q}(A_4^*/A_0^*) = 1.2384$
$\bar{q}(A_5^*/A_0^*) = 1.0168$

$f = 3/2$
$\bar{q}(A_1^*/A_0^*) = 2.4037$
$\bar{q}(A_2^*/A_0^*) = 2.3894$
$\bar{q}(A_3^*/A_0^*) = 1.1920$
$\bar{q}(A_4^*/A_0^*) = 0.9570$
$\bar{q}(A_5^*/A_0^*) = 0.9546$

$f = 1$
$\bar{q}(A_1^*/A_0^*) = 4.7382$
$\bar{q}(A_2^*/A_0^*) = 4.7291$
$\bar{q}(A_3^*/A_0^*) = 1.0840$
$\bar{q}(A_4^*/A_0^*) = 1.1075$
$\bar{q}(A_5^*/A_0^*) = 0.9431$

$f = 2/3$
$\bar{q}(A_1^*/A_0^*) = 6.0337$
$\bar{q}(A_2^*/A_0^*) = 5.8594$
$\bar{q}(A_3^*/A_0^*) = 0.9024$
$\bar{q}(A_4^*/A_0^*) = 1.1235$
$\bar{q}(A_5^*/A_0^*) = 0.8931$

$f = 1/3$
$\bar{q}(A_1^*/A_0^*) = 8.7061$
$\bar{q}(A_2^*/A_0^*) = 8.3326$
$\bar{q}(A_3^*/A_0^*) = 0.9447$
$\bar{q}(A_4^*/A_0^*) = 0.9775$
$\bar{q}(A_5^*/A_0^*) = 0.9133$

$f = 1/6$
$\bar{q}(A_1^*/A_0^*) = 16.638$
$\bar{q}(A_2^*/A_0^*) = 16.679$
$\bar{q}(A_3^*/A_0^*) = 1.1344$
$\bar{q}(A_4^*/A_0^*) = 1.1568$
$\bar{q}(A_5^*/A_0^*) = 1.0644$

$f = 1/12$
$\bar{q}(A_1^*/A_0^*) = 30.328$
$\bar{q}(A_2^*/A_0^*) = 30.254$
$\bar{q}(A_3^*/A_0^*) = 0.9977$
$\bar{q}(A_4^*/A_0^*) = 1.0030$
$\bar{q}(A_5^*/A_0^*) = 0.9703$

$f = 1/24$
$\bar{q}(A_1^*/A_0^*) = 66.229$
$\bar{q}(A_2^*/A_0^*) = 65.883$
$\bar{q}(A_3^*/A_0^*) = 1.3882$
$\bar{q}(A_4^*/A_0^*) = 1.1096$
$\bar{q}(A_5^*/A_0^*) = 1.4679$

$f = 1/48$
$\bar{q}(A_1^*/A_0^*) = 128.81$
$\bar{q}(A_2^*/A_0^*) = 129.02$
$\bar{q}(A_3^*/A_0^*) = 1.1277$
$\bar{q}(A_4^*/A_0^*) = 0.9591$
$\bar{q}(A_5^*/A_0^*) = 0.9492$

Neste caso temos $\hat{\alpha} = 0.8445$ e $\hat{\gamma} = 0.8857$ para $\bar{q}(A_1^*/A_0^*)$, fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 6.9904 \times (2 \times f)^{-0.8857}$$

e para $\bar{q}(A_2^*/A_0^*)$ temos $\hat{\alpha} = 0.8394$ e $\hat{\gamma} = 0.8882$, fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 6.9085 \times (2 \times f)^{-0.8882}$$

Nos três modelos o valor de 10^{α} é aproximadamente 7, tanto para $\bar{q}(A_1^*/A_0^*)$ como para $\bar{q}(A_2^*/A_0^*)$, e bem maior do que os observados nos outros modelos. Podemos suspeitar que os desvios-padrão diferentes das variáveis estão influenciando essa estimativa. Por outro lado, a mudança na estimativa de γ é bem pequena, sendo que as menores estimativas de γ, 0.8857 para $\bar{q}(A_1^*/A_0^*)$ e 0.8882 para $\bar{q}(A_2^*/A_0^*)$, são observados no modelo onde o desvio-padrão de ϵ_1 é 2.

3.4.3 Modelos com Três Variáveis Colineares

Após verificar o comportamento do diagnóstico proposto nos modelos onde tinhamos duas variáveis colineares, achamos que seria bastante interessante verificar a sensibilidade do diagnóstico na presença de três variáveis colineares.

Como nos modelos com duas variáveis colineares, apresentados anteriormente, os conjuntos de dados foram gerados através de simulações Monte Carlo, segundo o modelo $\mathbf{y} = \mathbf{X}\beta + \epsilon$, com cinco variáveis $(X_1, X_2, X_3, X_4, X_5)$ e 50 observações cada uma, onde as variáveis X_1, X_2 e X_3 são colineares.

Foram simulados dois tipos de modelos. O primeiro onde as variáveis X_1, X_4 e X_5 tem distribuição $N(0,1)$ e as variáveis X_2 e X_3 foram gerados como:

$$X_2 = X_1 + f_1\epsilon_1$$
$$X_3 = X_2 + f_2\epsilon_2$$

onde ϵ_1 e $\epsilon_2 \sim N(0,1)$ e f_1 e $f_2 \in \left(3, \frac{3}{2}, 1, \frac{2}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{12}, \frac{1}{24}, \frac{1}{48}\right)$. A princípio construímos 9 modelos considerando $f_1 = f_2$ e depois 8 modelos onde $f_1 \neq f_2$.

O segundo conjunto de modelos manteve a mesma estrutura do anterior sendo que agora $X_1 \sim N(0, 5.5^2), X_4 \sim N(0, 0.7^2)$ e $X_5 \sim N(0, 1.8^2)$. A justificativa para alterar o desvio-padrão das variáveis, está no fato de que em algumas situações, a alta variabilidade das variáveis X_i, em particular da variável construtora da colinearidade, afeta o valor da razão $\bar{q}$, para as variáveis envolvidas em colinearidade. Nos modelos onde tinhamos duas variáveis colineares, os desvios-padrão diferentes facilitaram a detecção da colinearidade. Achamos interessante verficar o que acontece na situação onde temos três variáveis colineares. Os resultados encontrados são bem interessantes e podem ser vistos nas tabelas a seguir.

Modelo com grau de colinearidade crescente e fatores iguais para X_2 e X_3

Analisando a tabela 3.28 vemos que já no fator $f = 3$, a presença da colinearidade entre X_2 e X_3 se manifesta, fazendo com que a razão $\bar{q}(A_i^*/A_0^*)$, $i = 2$ e 3, seja aproximadamente o inverso do fator de colinearidade, entretanto não sabemos qual a terceira variável envolvida. A partir de $f = \frac{2}{3}$ vai ficando mais evidente a participação da variável X_1. Para as variáveis X_4 e X_5, que não estão envolvidas em colinearidade, a razão, se mantém próxima de 1. Podemos ver também que o valor de $\bar{q}(A_1^*/A_0^*)$ é sempre maior que os valores de $\bar{q}(A_2^*/A_0^*)$ e $\bar{q}(A_3^*/A_0^*)$. Uma possível explicação está baseada no fato de que a variável X_1 carrega o peso da dependência das variáveis X_2 e X_3.

Tabela 3.28

Modelo supondo colinearidade entre X_1, X_2 e X_3,
sendo X_1, X_4 e $X_5 \sim N(0,1)$, ϵ, ϵ_1 e $\epsilon_2 \sim N(0,1)$.

$X_2 = X_1 + 3\epsilon_1$ $X_3 = X_1 + 3\epsilon_2$	$X_2 = X_1 + \frac{3}{2}\epsilon_1$ $X_3 = X_1 + \frac{3}{2}\epsilon_2$	$X_2 = X_1 + 1\epsilon_1$ $X_3 = X_1 + 1\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 1.0719$	$\bar{q}(A_1^*/A_0^*) = 1.6700$	$\bar{q}(A_1^*/A_0^*) = 1.7162$
$\bar{q}(A_2^*/A_0^*) = 0.3465$	$\bar{q}(A_2^*/A_0^*) = 0.5960$	$\bar{q}(A_2^*/A_0^*) = 1.0433$
$\bar{q}(A_3^*/A_0^*) = 0.3441$	$\bar{q}(A_3^*/A_0^*) = 0.8183$	$\bar{q}(A_3^*/A_0^*) = 1.3409$
$\bar{q}(A_4^*/A_0^*) = 1.2131$	$\bar{q}(A_4^*/A_0^*) = 1.2402$	$\bar{q}(A_4^*/A_0^*) = 0.9861$
$\bar{q}(A_5^*/A_0^*) = 1.0592$	$\bar{q}(A_5^*/A_0^*) = 0.9987$	$\bar{q}(A_5^*/A_0^*) = 1.6961$

$X_2 = X_1 + \frac{2}{3}\epsilon_1$ $X_3 = X_1 + \frac{2}{3}\epsilon_2$	$X_2 = X_1 + \frac{1}{3}\epsilon_1$ $X_3 = X_1 + \frac{1}{3}\epsilon_2$	$X_2 = X_1 + \frac{1}{6}\epsilon_1$ $X_3 = X_1 + \frac{1}{6}\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 1.9211$	$\bar{q}(A_1^*/A_0^*) = 3.6037$	$\bar{q}(A_1^*/A_0^*) = 8.6339$
$\bar{q}(A_2^*/A_0^*) = 1.3068$	$\bar{q}(A_2^*/A_0^*) = 2.5506$	$\bar{q}(A_2^*/A_0^*) = 6.2008$
$\bar{q}(A_3^*/A_0^*) = 1.4823$	$\bar{q}(A_3^*/A_0^*) = 2.6029$	$\bar{q}(A_3^*/A_0^*) = 7.5832$
$\bar{q}(A_4^*/A_0^*) = 0.9618$	$\bar{q}(A_4^*/A_0^*) = 0.8888$	$\bar{q}(A_4^*/A_0^*) = 0.8448$
$\bar{q}(A_5^*/A_0^*) = 0.8604$	$\bar{q}(A_5^*/A_0^*) = 0.8848$	$\bar{q}(A_5^*/A_0^*) = 0.7890$

$X_2 = X_1 + \frac{1}{12}\epsilon_1$	$X_2 = X_1 + \frac{1}{24}\epsilon_1$	$X_2 = X_1 + \frac{1}{48}\epsilon_1$
$X_3 = X_1 + \frac{1}{12}\epsilon_2$	$X_3 = X_1 + \frac{1}{24}\epsilon_2$	$X_3 = X_1 + \frac{1}{48}\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 20.472$	$\bar{q}(A_1^*/A_0^*) = 37.729$	$\bar{q}(A_1^*/A_0^*) = 72.395$
$\bar{q}(A_2^*/A_0^*) = 15.706$	$\bar{q}(A_2^*/A_0^*) = 26.143$	$\bar{q}(A_2^*/A_0^*) = 40.466$
$\bar{q}(A_3^*/A_0^*) = 10.961$	$\bar{q}(A_3^*/A_0^*) = 22.909$	$\bar{q}(A_3^*/A_0^*) = 50.091$
$\bar{q}(A_4^*/A_0^*) = 0.7174$	$\bar{q}(A_4^*/A_0^*) = 1.1111$	$\bar{q}(A_4^*/A_0^*) = 1.0228$
$\bar{q}(A_5^*/A_0^*) = 1.0274$	$\bar{q}(A_5^*/A_0^*) = 0.7423$	$\bar{q}(A_5^*/A_0^*) = 0.9100$

Como já haviamos feito nos modelos com duas variáveis colineares, tentamos escrever $\bar{q}(A_i^*/A_0^*)$ $i = 1, 2$ e 3 como uma função do fator de colinearidade através da equação de regressão,

$$\log \bar{q} = \alpha - \gamma \log(f)$$

Neste caso temos $\hat{\alpha} = 0.2899$ e $\hat{\gamma} = 0.8950$ para $\bar{q}(A_1^*/A_0^*)$ fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 1.9494 \times (1 \times f)^{-0.8950}$$

Os valores estimados pela equação de regressão e seu respectivo erro relativo, dado pela equação (3.7), podem ser vistos na tabela abaixo:

Tabela 3.29

Fator	**Obs.**	**Est.**	$\mid \Delta\% \mid$
3	1.0719	0.7293	46.95
$\frac{3}{2}$	1.6700	1.3561	23.147
1	1.7162	1.9494	21.963
$\frac{2}{3}$	1.9211	2.8022	88.110
$\frac{1}{3}$	3.6037	5.2110	30.844
$\frac{1}{6}$	8.6339	9.6905	10.908
$\frac{1}{12}$	20.472	18.021	38.910
$\frac{1}{24}$	37.729	33.511	12.586
$\frac{1}{48}$	72.395	62.318	16.171

e a média dos erros é:

$$\frac{1}{9}\sum \mid \Delta\% \mid = 31.06$$

Para $\bar{q}(A_2^*/A_0^*)$ temos $\hat{\alpha} = 0.0425$ e $\hat{\gamma} = 0.9978$ fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 1.1028 \times (1 \times f)^{-0.9978}$$

Tabela 3.30

Fator	Obs.	Est.	$\mid \Delta\% \mid$
3	0.3465	0.3685	6.2144
$\frac{3}{2}$	0.5960	0.7359	19.011
1	1.0433	1.1028	5.3954
$\frac{2}{3}$	1.3068	1.6527	20.929
$\frac{1}{3}$	2.5506	3.3004	22.3718
$\frac{1}{6}$	6.2008	6.5908	5.9173
$\frac{1}{12}$	15.706	13.162	19.333
$\frac{1}{24}$	26.143	26.283	0.5319
$\frac{1}{48}$	40.466	52.485	22.901

e a média dos erros é:

$$\frac{1}{9}\sum \mid \Delta\% \mid = 13.66$$

e para $\bar{q}(A_3^*/A_0^*)$ que $\hat{\alpha} = 0.0173$ e $\hat{\gamma} = 0.9852$ fazendo com que a equação (3.9) seja, numericamente, dada por:

$$q = 1.0406 \times (1 \times f)^{-0.9852}$$

Tabela 3.31

Fator	Obs.	Est.	$\mid \Delta\% \mid$
3	0.3441	0.3526	2.4107
$\frac{3}{2}$	0.8183	0.6979	17.252
1	1.3409	1.0406	28.858
$\frac{2}{3}$	1.4823	1.5516	4.4664
$\frac{1}{3}$	2.6029	3.0715	15.256
$\frac{1}{6}$	7.5832	6.0802	24.719
$\frac{1}{12}$	10.961	12.036	8.934
$\frac{1}{24}$	22.909	23.827	3.8524
$\frac{1}{48}$	55.091	47.167	6.1981

e a média dos erros é:

$$\frac{1}{9}\sum \mid \Delta\% \mid = 12.44$$

Observamos que os valores de 10^{α} e γ para a equação de regressão ajustada para os valores de $\bar{q}(A_1^*/A_0^*)$, estão um pouco mais distantes de 1 do que os demais. Como já haviamos dito anteriormente em relação aos modelos com duas variáveis colineares, este fato pode estar ocorrendo pois a variável X_1, sendo a variável formadora das outras duas, carrega o peso maior da colinearidade. Podemos observar também que a maior média dos erros está associada a variável X_1.

Neste modelo consideramos os fatores de colinearidade iguais para as variáveis X_2 e X_3, a seguir vamos considerar fatores de colinearidade diferentes para as duas variáveis e ver como a razão $\bar{q}$ se comporta.

Modelo com grau de colinearidade crescente e fatores diferentes para X_2 e X_3

Analisando a tabela 3.32 vemos que para fatores diferentes, não é tão simples identificar quais são as três variáveis envolvidas em colinearidade. Por exemplo, quando os fatores

são $f_1 = 3$ e $f_2 = \frac{3}{2}$ conseguimos identificar que X_2 e X_3 estão envolvidas em colinearidade entretanto não sabemos qual é a terceira variável envolvida. A identificação das três variáveis envolvidas vai ficando mais clara a medida que o grau de colinearidade aumenta.

Tabela 3.32

Modelo supondo colinearidade entre X_1, X_2 e X_3,
sendo X_1, X_4 e $X_5 \sim N(0,1)$, ϵ, ϵ_1 e $\epsilon_2 \sim N(0,1)$.

$X_2 = X_1 + 3\epsilon_1$ $X_3 = X_1 + \frac{3}{2}\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 1.3790$
$\bar{q}(A_2^*/A_0^*) = 0.5041$
$\bar{q}(A_3^*/A_0^*) = 0.6287$
$\bar{q}(A_4^*/A_0^*) = 0.9077$
$\bar{q}(A_5^*/A_0^*) = 1.0531$

$X_2 = X_1 + \frac{3}{2}\epsilon_1$ $X_3 = X_1 + 3\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 1.0247$
$\bar{q}(A_2^*/A_0^*) = 0.7571$
$\bar{q}(A_3^*/A_0^*) = 0.2542$
$\bar{q}(A_4^*/A_0^*) = 0.7633$
$\bar{q}(A_5^*/A_0^*) = 0.8081$

$X_2 = X_1 + 1\epsilon_1$ $X_3 = X_1 + \frac{1}{6}\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 7.5148$
$\bar{q}(A_2^*/A_0^*) = 0.8344$
$\bar{q}(A_3^*/A_0^*) = 6.8089$
$\bar{q}(A_4^*/A_0^*) = 1.1064$
$\bar{q}(A_5^*/A_0^*) = 1.1120$

$X_2 = X_1 + \frac{2}{3}\epsilon_1$ $X_3 = X_1 + 1\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 2.4900$
$\bar{q}(A_2^*/A_0^*) = 1.7063$
$\bar{q}(A_3^*/A_0^*) = 0.9454$
$\bar{q}(A_4^*/A_0^*) = 1.5907$
$\bar{q}(A_5^*/A_0^*) = 1.0734$

$X_2 = X_1 + \frac{1}{3}\epsilon_1$ $X_3 = X_1 + \frac{1}{12}\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 13.722$
$\bar{q}(A_2^*/A_0^*) = 4.8656$
$\bar{q}(A_3^*/A_0^*) = 11.467$
$\bar{q}(A_4^*/A_0^*) = 0.9822$
$\bar{q}(A_5^*/A_0^*) = 1.0151$

$X_2 = X_1 + \frac{1}{6}\epsilon_1$ $X_3 = X_1 + \frac{2}{3}\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 10.996$
$\bar{q}(A_2^*/A_0^*) = 11.051$
$\bar{q}(A_3^*/A_0^*) = 1.9403$
$\bar{q}(A_4^*/A_0^*) = 0.9492$
$\bar{q}(A_5^*/A_0^*) = 1.2577$

$X_2 = X_1 + \frac{1}{12}\epsilon_1$ $X_3 = X_1 + \frac{1}{24}\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 20.188$
$\bar{q}(A_2^*/A_0^*) = 13.069$
$\bar{q}(A_3^*/A_0^*) = 18.692$
$\bar{q}(A_4^*/A_0^*) = 1.0863$
$\bar{q}(A_5^*/A_0^*) = 1.0147$

$X_2 = X_1 + \frac{1}{24}\epsilon_1$ $X_3 = X_1 + \frac{1}{3}\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 27.340$
$\bar{q}(A_2^*/A_0^*) = 25.673$
$\bar{q}(A_3^*/A_0^*) = 3.5772$
$\bar{q}(A_4^*/A_0^*) = 1.0230$
$\bar{q}(A_5^*/A_0^*) = 0.9164$

Modelos com grau de colinearidade crescente onde as variáveis X_1, X_4 e X_5 tem desvios-padrão diferentes e fatores iguais para X_2 e X_3

Analisando a tabela 3.33 vemos que já no primeiro fator, $f = 3$, o valor da razão $\bar{q}$ mostra quais são as três variáveis envolvidas em colinearidade. Já haviamos visto nos modelos com duas variáveis colineares que o fato das variáveis terem desvios-padrão diferentes, aumenta o valor da razão $\bar{q}$ para as variáveis envolvidas em colinearidade, facilitando a sua identificação. Entretanto nada acontece com o patamar, próximo de 1, apresentado pelas demais variáveis, isto é, esse patamar só se altera se o desvio de ϵ_1 for modificado.

Tabela 3.33

Modelo supondo colinearidade entre X_1, X_2 e X_3,
sendo $X_1 \sim N(0, 5.5^2)$, $X_4 \sim N(0, 0.7^2)$ e $X_5 \sim N(0, 1.8^2)$,
ϵ, ϵ_1 e $\epsilon_2 \sim N(0, 1)$.

$X_2 = X_1 + 3\epsilon_1$ $X_3 = X_1 + 3\epsilon_2$	$X_2 = X_1 + \frac{3}{2}\epsilon_1$ $X_3 = X_1 + \frac{3}{2}\epsilon_2$	$X_2 = X_1 + 1\epsilon_1$ $X_3 = X_1 + 1\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 2.7232$	$\bar{q}(A_1^*/A_0^*) = 5.5365$	$\bar{q}(A_1^*/A_0^*) = 1.2922$
$\bar{q}(A_2^*/A_0^*) = 2.4517$	$\bar{q}(A_2^*/A_0^*) = 3.8321$	$\bar{q}(A_2^*/A_0^*) = 1.1038$
$\bar{q}(A_3^*/A_0^*) = 2.0541$	$\bar{q}(A_3^*/A_0^*) = 4.2059$	$\bar{q}(A_3^*/A_0^*) = 1.1001$
$\bar{q}(A_4^*/A_0^*) = 0.9104$	$\bar{q}(A_4^*/A_0^*) = 1.0291$	$\bar{q}(A_4^*/A_0^*) = 0.7107$
$\bar{q}(A_5^*/A_0^*) = 0.8796$	$\bar{q}(A_5^*/A_0^*) = 0.9465$	$\bar{q}(A_5^*/A_0^*) = 0.8344$

A tabela continua na página seguinte

$X_2 = X_1 + \frac{2}{3}\epsilon_1$	$X_2 = X_1 + \frac{1}{3}\epsilon_1$	$X_2 = X_1 + \frac{1}{6}\epsilon_1$
$X_3 = X_1 + \frac{2}{3}\epsilon_2$	$X_3 = X_1 + \frac{1}{3}\epsilon_2$	$X_3 = X_1 + \frac{1}{6}\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 10.905$	$\bar{q}(A_1^*/A_0^*) = 15.450$	$\bar{q}(A_1^*/A_0^*) = 47.288$
$\bar{q}(A_2^*/A_0^*) = 7.0094$	$\bar{q}(A_2^*/A_0^*) = 12.442$	$\bar{q}(A_2^*/A_0^*) = 33.107$
$\bar{q}(A_3^*/A_0^*) = 5.6648$	$\bar{q}(A_3^*/A_0^*) = 12.763$	$\bar{q}(A_3^*/A_0^*) = 31.001$
$\bar{q}(A_4^*/A_0^*) = 0.8674$	$\bar{q}(A_4^*/A_0^*) = 0.7927$	$\bar{q}(A_4^*/A_0^*) = 1.0528$
$\bar{q}(A_5^*/A_0^*) = 1.0651$	$\bar{q}(A_5^*/A_0^*) = 0.7839$	$\bar{q}(A_5^*/A_0^*) = 1.1612$

$X_2 = X_1 + \frac{1}{12}\epsilon_1$	$X_2 = X_1 + \frac{1}{24}\epsilon_1$	$X_2 = X_1 + \frac{1}{48}\epsilon_1$
$X_3 = X_1 + \frac{1}{12}\epsilon_2$	$X_3 = X_1 + \frac{1}{24}\epsilon_2$	$X_3 = X_1 + \frac{1}{48}\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 80.752$	$\bar{q}(A_1^*/A_0^*) = 189.77$	$\bar{q}(A_1^*/A_0^*) = 390.56$
$\bar{q}(A_2^*/A_0^*) = 51.806$	$\bar{q}(A_2^*/A_0^*) = 163.39$	$\bar{q}(A_2^*/A_0^*) = 313.98$
$\bar{q}(A_3^*/A_0^*) = 61.158$	$\bar{q}(A_3^*/A_0^*) = 126.64$	$\bar{q}(A_3^*/A_0^*) = 246.52$
$\bar{q}(A_4^*/A_0^*) = 0.9896$	$\bar{q}(A_4^*/A_0^*) = 1.2820$	$\bar{q}(A_4^*/A_0^*) = 0.8884$
$\bar{q}(A_5^*/A_0^*) = 0.7547$	$\bar{q}(A_5^*/A_0^*) = 1.5997$	$\bar{q}(A_5^*/A_0^*) = 0..9018$

Como já foi visto nos modelos com duas variáveis colineares, com desvios-padrão diferentes para as variáveis, o valor de 10^{α} fica bem diferente de 1. Este fato era esperado pois se observarmos os valores de $\bar{q}(A_i^*/A_0^*)$ $i = 1, 2$ e 3 na tabela 3.33, vemos claramente que eles são superiores aos observados no modelo onde as variáveis apresentam desvios-padrão iguais, tabela 3.28. Observamos também que o maior erro relativo, associado ao ajuste das equações de regressão, é observado para $\bar{q}(A_1^*/A_0^*)$, que está relacionado a variável X_1 que é a construtora da colinearidade. As equações de regressão ajustadas para $\bar{q}(A_i^*/A_0^*)$ $i = 1, 2$ e 3, os valores estimados e seu respectivo erro relativo, dado pela equação (3.7), podem ser vistos a seguir.

Neste caso temos $\hat{\alpha} = 0.7324$ e $\hat{\gamma} = 1.0930$ para $\bar{q}(A_1^*/A_0^*)$ fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 5.4001 \times (1 \times f)^{-1.0930}$$

Os valores estimados podem ser vistos na tabela abaixo:

Tabela 3.34

Fator	**Obs.**	**Est.**	$\mid \Delta\% \mid$
3	2.7232	1.6252	67.561
$\frac{3}{2}$	5.5365	3.4668	59.701
1	1.2922	5.4001	76.071
$\frac{2}{3}$	10.905	8.4114	29.645
$\frac{1}{3}$	15.450	17.943	13.894
$\frac{1}{6}$	47.288	38.275	23.546
$\frac{1}{12}$	80.752	81.648	1.0978
$\frac{1}{24}$	189.77	174.17	8.9570
$\frac{1}{48}$	390.56	371.53	5.1210

e a média dos erros é:

$$\frac{1}{9}\sum \mid \Delta\% \mid = 31.73$$

Para $\bar{q}(A_2^*/A_0^*)$ temos $\hat{\alpha} = 0.6713$ e $\hat{\gamma} = 1.0670$, fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 4.6914 \times (1 \times f)^{-1.0670}$$

Os valores estimados podem ser vistos na tabela abaixo:

Tabela 3.35

Fator	**Obs.**	**Est.**	$\mid \Delta\% \mid$
3	2.4517	1.4528	68.757
$\frac{3}{2}$	3.8321	3.0438	25.898
1	1.1038	4.6914	76.472
$\frac{2}{3}$	7.0094	7.2309	3.063
$\frac{1}{3}$	12.442	15.149	17.870
$\frac{1}{6}$	33.107	31.739	4.311
$\frac{1}{12}$	51.806	66.495	22.090
$\frac{1}{24}$	163.39	139.31	17.284
$\frac{1}{48}$	313.98	291.87	7.5759

e a média dos erros é:

$$\frac{1}{9}\sum \mid \Delta\% \mid = 27.04$$

e para $\bar{q}(A_3^*/A_0^*)$ temos $\hat{\alpha} = 0.6637$ e $\hat{\gamma} = 1.0121$, fazendo com que a equação (3.9) seja, numericamente, dada por:

$$\bar{q} = 4.61 \times (1 \times f)^{-1.0121}$$

Os valores estimados podem ser vistos na tabela abaixo:

Tabela 3.36

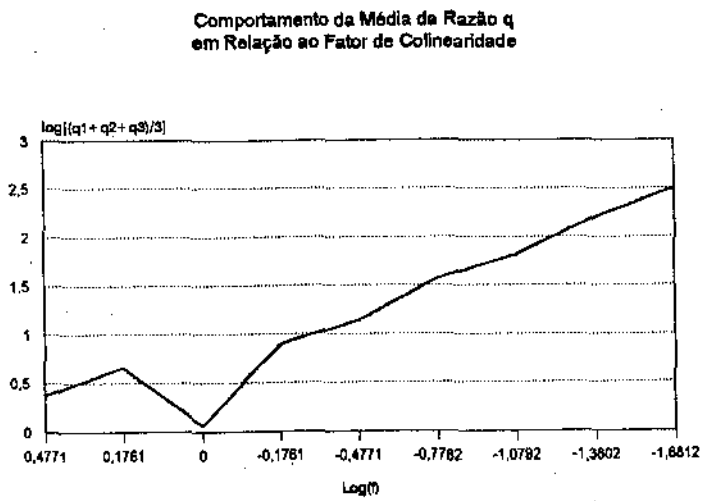
Fator	**Obs.**	**Est.**	$\mid \Delta\% \mid$
3	2.0541	1.5164	35.189
$\frac{3}{2}$	4.2059	3.0583	37.524
1	1.1001	4.6100	76.137
$\frac{2}{3}$	5.6648	6.9490	18.480
$\frac{1}{3}$	12.763	14.015	8.9339
$\frac{1}{6}$	31.001	28.266	3.8801
$\frac{1}{12}$	61.158	57.009	7.2786
$\frac{1}{24}$	126.64	174.98	10.143
$\frac{1}{48}$	246.52	231.89	6.3082

e a média dos erros é:

$$\frac{1}{9}\sum \mid \Delta\% \mid = 22.65$$

Poderia ser questionado que quando o fator de colinearidade é igual a 1, todas as razões são aproximadamente iguais a 1, impossibilitando a identificação das variáveis colineares. Realmente, neste caso a igualdade entre o fator de colinearidade e o desvio de ϵ_1 faz com que a identificação fique prejudicada para esse único fator. O grafico 4 mostra o comportamento desse e dos demais pontos.

Gráfico 4



Modelos com grau de colinearidade crescente onde as variáveis X_1, X_4 e X_5 tem desvios e fatores diferentes para X_2 e X_3

A tabela 3.37 nos mostra que mesmo tendo fatores de colinearidade diferentes, o fato das variáveis terem desvios-padrão diferentes facilita a detecção da colinearidade. Como no modelo anterior a identificação das variáveis envolvidas em colinearidade pode ser feita logo para os primeiro fatores de colinearidade. O valor da razão $\bar{q}(A_i^*/A_0^*)$ $i = 4$ e 5 continua inalterado, isto é, próximo de 1.

Tabela 3.37

Modelo supondo colinearidade entre X_1, X_2 e X_3,
sendo $X_1 \sim N(0, 5.5^2)$, $X_4 \sim N(0, 0.7^2)$ e $X_5 \sim N(0, 1.8^2)$,
ϵ, ϵ_1 e $\epsilon_2 \sim N(0, 1)$.

$X_2 = X_1 + 3\epsilon_1$ $X_3 = X_1 + \frac{3}{2}\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 2.6270$
$\bar{q}(A_2^*/A_0^*) = 1.6035$
$\bar{q}(A_3^*/A_0^*) = 2.4054$
$\bar{q}(A_4^*/A_0^*) = 0.9983$
$\bar{q}(A_5^*/A_0^*) = 0.9850$

$X_2 = X_1 + \frac{3}{2}\epsilon_1$ $X_3 = X_1 + 3\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 3.9263$
$\bar{q}(A_2^*/A_0^*) = 3.8166$
$\bar{q}(A_3^*/A_0^*) = 2.0959$
$\bar{q}(A_4^*/A_0^*) = 1.0095$
$\bar{q}(A_5^*/A_0^*) = 0.9631$

$X_2 = X_1 + 1\epsilon_1$ $X_3 = X_1 + \frac{1}{6}\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 33.809$
$\bar{q}(A_2^*/A_0^*) = 4.9767$
$\bar{q}(A_3^*/A_0^*) = 33.713$
$\bar{q}(A_4^*/A_0^*) = 0.9220$
$\bar{q}(A_5^*/A_0^*) = 1.1263$

$X_2 = X_1 + \frac{2}{3}\epsilon_1$ $X_3 = X_1 + 1\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 8.8999$
$\bar{q}(A_2^*/A_0^*) = 7.7960$
$\bar{q}(A_3^*/A_0^*) = 5.4793$
$\bar{q}(A_4^*/A_0^*) = 0.8684$
$\bar{q}(A_5^*/A_0^*) = 1.1427$

$X_2 = X_1 + \frac{1}{3}\epsilon_1$ $X_3 = X_1 + \frac{1}{12}\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 60.314$
$\bar{q}(A_2^*/A_0^*) = 21.692$
$\bar{q}(A_3^*/A_0^*) = 51.789$
$\bar{q}(A_4^*/A_0^*) = 1.0157$
$\bar{q}(A_5^*/A_0^*) = 1.1860$

$X_2 = X_1 + \frac{1}{6}\epsilon_1$ $X_3 = X_1 + \frac{2}{3}\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 27.063$
$\bar{q}(A_2^*/A_0^*) = 27.669$
$\bar{q}(A_3^*/A_0^*) = 9.0888$
$\bar{q}(A_4^*/A_0^*) = 0.9386$
$\bar{q}(A_5^*/A_0^*) = 0.9941$

$X_2 = X_1 + \frac{1}{12}\epsilon_1$ $X_3 = X_1 + \frac{1}{24}\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 127.4680$
$\bar{q}(A_2^*/A_0^*) = 67.2828$
$\bar{q}(A_3^*/A_0^*) = 96.4651$
$\bar{q}(A_4^*/A_0^*) = 1.1093$
$\bar{q}(A_5^*/A_0^*) = 0.9008$

$X_2 = X_1 + \frac{1}{24}\epsilon_1$ $X_3 = X_1 + \frac{1}{3}\epsilon_2$
$\bar{q}(A_1^*/A_0^*) = 165.5766$
$\bar{q}(A_2^*/A_0^*) = 170.6473$
$\bar{q}(A_3^*/A_0^*) = 20.0456$
$\bar{q}(A_4^*/A_0^*) = 1.1909$
$\bar{q}(A_5^*/A_0^*) = 1.0655$

3.5 Conclusões

Um dos objetivos do nosso trabalho foi fazer uma revisão dos métodos mais utilizados para diagnosticar a colinearidade. Vimos que o método desenvolvido por BKW é atualmente o mais utilizado e também o mais eficiente entre os estudados. Entretanto, pudemos perceber, ao avaliarmos cuidadosamente o diagnóstico, que ainda existem algumas deficiências. As recomendações dadas por BKW fazem com que a detecção da colinearidade só seja feita nos casos onde o coeficiente de correlação entre as variáveis é muito alta, $> 90\%$. Um outro ponto, está na utilização conjunta dos índices de condição e da decomposição da variância. Como os índices de condição só conseguem mostrar que existem variáveis colineares no modelo sem identificá-las, sendo sempre necessária a utilização da decomposição da variância para mostrar quais as variáveis envolvidas em colinearidade, levantamos então a seguinte questão: Será que os índices de condição são realmente necessários? Sob o nosso ponto de vista achamos que sua utilização é dispensável, a proporção de decomposição da variância sozinha é capaz de mostrar a existência da colinearidade e também identificar as variáveis envolvidas.

Essa pequena falha do diagnóstico proposto por BKW, nos levou a pensar na possibilidade de introduzir algum outro método de diagnosticar a colinearidade, surgindo então outra questão: Será que a aplicação do bootstrap nos diagnósticos nos daria alguma informação adicional relativa à colinearidade? Vimos que apesar da ampla utilização do método bootstrap em diversas situações, não encontramos na literatura a sua utilização em diagnósticos de colinearidade. Estava aberta então uma lacuna, onde o nosso trabalho pode ser visto como um começo, mas apenas o começo do preenchimento deste espaço.

Apresentamos quatro diagnósticos que utilizam o bootstrap como uma ferramenta na detecção da colinearidade: o coeficiente de variação estimado, os índices de condição, a proporção de decomposição da variância e a razão $\bar{q}$, todos calculados a partir de reamostragens feitas no par $(\mathbf{y}, \mathbf{X})$. Construimos a princípio um modelo básico com cinco variáveis, onde duas delas apresentavam colinearidade e avaliamos a sensibilidade dos diagnósticos.

O coeficiente de variação estimado consegue identificar as duas variáveis envolvidas em colinearidade, mas apesar da variável construtora (da colinearidade) ser identificada rapidamente, a segunda variável envolvida só é identificada quando o grau de colinearidade é alto.

O comportamento dos índices de condição e da proporção de decomposição da variância é análogo ao caso onde eles são calculados na matriz de dados original, **X**. Eles continuam detectando somente graus de colinearidade altos entre as variáveis.

Finalmente, a análise da razão $\bar{q}$ mostrou ser um instrumento analítico poderoso na detecção da colinearidade, sendo capaz de identificar, no modelo básico, as variáveis envolvidas em colinearidade e quantificar o grau de dependência entre elas, mesmo quando a colinearidade ainda é moderada.

A partir do modelo básico tivemos uma indicação de que a razão $\bar{q}$ seria um diagnóstico poderoso na detecção da colinearidade. Foram simulados vários modelos com duas e três variáveis colineares, onde ainda foram feitas modificações, nos desvios-padrão das variáveis e do ϵ_1, supondo sempre a construção $X_2 = X_1 + f\epsilon_1$. O nosso objetivo com as simulações foi avaliar o comportamento da razão $\bar{q}$ em diversas situações onde a colinearidade estava presente. Pudemos confirmar o que havia sido visto no modelo básico, a razão $\bar{q}$ é um diagnóstico sensível a presença da colinearidade.

No caso onde as variáveis tem distribuição $N(0,1)$ e a colinearidade envolve duas variáveis a razão $\bar{q}$ é capaz de identificar as variáveis envolvidas e ainda através das simulações dar um indicativo do grau de dependência entre elas, pois o valor da razão $\bar{q}$ para as variáveis envolvidas em colinearidade é aproximadamente o inverso do fator de colinearidade, sendo este fato manifestado primeiro na variável construida a partir da colinearidade. Entretanto, quando as variáveis tem distribuição $N(0,\sigma^2)$ onde $\sigma^2 \neq 1$, as variáveis envolvidas em colinearidade se destacam das demais já no primeiro fator de colinearidade, neste caso os valores da razão $\bar{q}$ não são mais aproximadamente o inverso do fator de colinearidade, mas através das simulações feitas podemos ter ainda um indicativo do grau de dependência entre as variáveis.

Quando a colinearidade envolve três variáveis a sensibilidade da razão $\bar{q}$ continua inalterada se as variáveis tem distribuição $N(0,\sigma^2)$ onde $\sigma^2 \neq 1$, a detecção da colinearidade é feita independente dos fatores de colinearidade serem iguais ou diferentes para as duas variáveis e independente do grau de colinearidade. Por outro lado, quando as variáveis tem distribuição $N(0,1)$, a detecção é um pouco mais difícil. Se os fatores de colinearidade são diferentes e misturam altos e baixos graus de colinearidade a detecção só é feita quando as duas variáveis tem um grau de colinearidade alto. Quando os fatores de colinearidade são iguais para as duas variáveis percebemos já no primeiro fator que temos duas variável envolvidas, mas só a partir de $f = \frac{2}{3}$ é que conseguimos identificar a terceira variável.

Mais um caso, que não foi tratado nas simulações, vai ser observado nas aplicações se a colinearidade envolve conjuntos de variáveis com graus de dependência diferentes, mas separado em blocos o diagnóstico também é capaz de identificar. Entretanto se as variáveis apresentam conjuntos de variáveis com colinearidade, mas onde o grau de envolvimento é similar em todos os casos, o diagnóstico sozinho só é capaz de identificar as variáveis envolvidas em colinearidade, mas não consegue separar os grupos. Neste caso é necessária a utilização da proporção de decomposição da variância para auxiliar a identificação.

Durante as simulações pudemos observar também alguns fatores que influênciam os valores da razão $\bar{q}$, o desvio-padrão de ϵ_1 alteram o valor de $\bar{q}(A_2^*/A_0^*)$, isto é, se compararmos ao modelo onde $\epsilon_1 \sim N(0,1)$, uma alteração feita no desvio-padrão de ϵ_1 é refletida em $\bar{q}(A_2^*/A_0^*)$ através de uma fator multiplicativo igual ao inverso do desvio-padrão de ϵ_1.

As simulações sugerem uma conjectura: Quando f é menor do que o desvio-padrão da variável construtora, no caso X_1, a presença da colinearidade pode ser detectada também nessa variável, a colinearidade nas variáveis construidas pode ser detectada em todos os fatores, menos no fator 1.

Temos também que se $\sigma_{\epsilon_1} > \sigma_{X_1}$ os valores de 10^α e γ ficam alterados, isto é, se afastam do valor 1 obtido na equação de regressão ajustada para $\bar{q}(A_i^*/A_0^*)$ $i = 1$ e 2, no

modelo básico, neste caso estamos considerando os modelos onde todas as variáveis tem o mesmo desvio-padrão. Quando as variáveis tem distribuição $N(0, \sigma^2)$ onde $\sigma^2 \neq 1$, o valor de 10^{α} fica bem distante de 1, isto é devido aos altos valores de $\bar{q}$ já nos primeiro fatores de colinearidade.

Nos preocupamos em variar o desvio-padrão de ϵ_1 e das variáveis para observar o comportamento da razão $\bar{q}$, mas não mexemos no desvio de ϵ do modelo total, ele foi sempre considerado como $N(0, 1)$, não sabemos então qual seria a sua influência no valor da razão $\bar{q}$.

As variáveis foram geradas sempre com distribuição Normal, outras distribuições devem ser exploradas para termos uma maior abrangência do diagnóstico.

A forma da colinearidade sempre ficou a mesma, isto é, $X_i = X_k + f\epsilon_i$. Outras formas, por exemplo, $X_i = \sum_{k=1}^{r} a_{l_k} X_{l_k} + f_i \epsilon_i$, onde $r \leq p - 1$, $l_k \neq i$, $k = 1, \ldots, r$ não foram exploradas. Esta forma pode ser importante no caso em que tem três ou mais variáveis envolvidas em colinearidade.

O resultados apresentados aqui são ainda experimentais, muito trabalho ainda tem que ser feito para uma avaliação completa da razão $\bar{q}$, inclusive todo o desenvolvimento teórico do que foi observado através das simulações.

Capítulo 4

Aplicações

As aplicações feitas neste capítulo tem como objetivo comparar os diagnósticos desenvolvidos por BKW e o diagnóstico proposto $\bar{q}$, em relação a sensibilidade em diagnosticar a presença da colinearidade. Utilizamos duas situações onde a colinearidade entre variáveis se apresenta de maneira diferente e analisamos a eficácia dos diagnósticos.

4.1 Aplicação 1: Bellyfat

Este exemplo foi retirado do Manual *SAS System for Regression* [21]. O objetivo deste estudo é estimar a porcentagem de gordura (FAT) na barriga de porcos. Como essa medida é determinada através de um procedimento muito caro, é interessante ver se este valor pode ser estimado através de outras medidas mais fáceis e mais baratas que são determinadas a partir da carcaça do porco. Dez variáveis foram medidas em uma amostra de quarenta e cinco carcaças de porcos. As variáveis são listadas à seguir:

AVBF: média de três medidas da espessura da gordura da parte traseira do porco.

MUS: score da musculatura da carcaça. Quanto maior o score, mais musculatura e menos gordura.

LEA: área de um corte do lombo.

DEP: média de três medidas da espessura da gordura oposta à décima costela.

LWT: peso da carcaça viva.

CWT: peso da carcaça morta.

WTWAT: é uma medida utilizada para determinar a gravidade específica.

DPSL: média de três medidas da profundidade da barriga.

LESL: média da medida de ausência de gordura em três cortes da barriga do porco.

BELWT: peso total da barriga.

Utilizando os diagnósticos propostos por BKW, o número e os índices de condição e a proporção de decomposição da variância, obtemos os seguintes resultados:

Tabela 4.1

Autovalores	**Índice de Condição**	**Prop. Var. AVBF**	**Prop. Var. MUS**	**Prop. Var. LEA**	**Prop. Var. DEP**
4.040977	1.000000	0.0065	0.0000	0.0059	0.0031
2.967376	1.166963	0.0000	0.0221	0.0005	0.0001
0.877602	2.145826	0.0156	0.0204	0.0065	0.0000
0.758552	2.308079	0.0893	0.0231	0.0118	0.0071
0.497983	2.848628	0.0046	0.0147	0.0053	0.0011
0.479097	2.904234	0.0358	0.3889	0.0182	0.0014
0.226938	4.219774	0.0732	0.0000	0.0395	0.0105
0.096617	6.467208	0.2471	0.1047	0.5933	0.0801
0.034525	10.8186	0.5030	0.3534	0.1144	0.7708
0.020333	14.0975	0.0250	0.0727	0.2046	0.1258

Obs: A tabela continua na página seguinte

Prop. Var. LWT	Prop. Var. CWT	Prop. Var. WTWAT	Prop. Var. DPSL	Prop. Var. LESL	Prop. Var. BELWT
0.0001	0.0001	0.0037	0.0102	0.0154	0.0004
0.0043	0.0038	0.0006	0.0015	0.0000	0.0264
0.0011	0.0024	0.0096	0.2232	0.0110	0.1253
0.0002	0.0012	0.0014	0.2387	0.0153	0.0034
0.0050	0.0012	0.0000	0.1429	0.2887	0.3062
0.0049	0.0048	0.0067	0.0085	0.0307	0.0422
0.0171	0.0077	0.0512	0.0848	0.5760	0.2711
0.0113	0.0157	0.1187	0.1071	0.0428	0.0753
0.0425	0.0001	0.8029	0.0351	0.0199	0.0248
0.9136	0.9629	0.0051	0.1480	0.0002	0.1249

Se seguissemos fielmente a recomendação de BKW, diriamos que estas variáveis não estão envolvidas em colinearidade, pois neste caso podemos observar que o maior índice de condição é igual a 14.1. E a recomendação dada por BKW diz que ao observarmos um número de condição > 30, que seria um indicativo de colinearidade, devemos então olhar a proporção de decomposição da variância para podermos descobrir quais as variáveis envolvidas em colinearidade. Neste caso, como o número de condição é inferior a 30, descartariamos a hipótese de colinearidade. Esta decisão seria um pouco precipitada, pois se observarmos as análises referentes ao ajuste do modelo linear percebemos que alguma coisa estranha está acontecendo.

Tabela 4.2

		Análise de	Variância		
Fonte de Variação	**Graus de Liberdade**	**Soma de Quadrados**	**Média de Quadrados**	**F** -	$p-valor$ -
Regressão	10	861.79	86.18	13.437	0.0001
Erro	34	218.06	6.41	-	-
Total	44	1079.85	-	-	-

Tabela 4.3

Variáveis	**g.l.**	**Parâmetros Estimados**	**Erro Padrão**	**t para H_0: Parâmetro=0**	$Prob > \|t\|$
Intercepto	1	24.8575	21.0138	1.183	0.2451
AVBF	1	-5.4725	3.6560	-1.497	0.1437
MUS	1	-0.5866	0.3072	-1.909	0.0647
LEA	1	-0.9893	1.7079	-0.579	0.5663
DEP	1	9.0649	4.8865	1.855	0.0723
LWT	1	0.1505	0.2047	0.735	0.4673
CWT	1	0.0304	0.2096	0.145	0.8857
WTWAT	1	-1.8125	2.9654	-0.611	0.5451
DPSL	1	-2.1419	3.5475	-0.604	0.5500
LESL	1	-0.4327	0.2899	-1.492	0.1448
BELWT	1	0.6867	0.4722	1.454	0.1550

Observando as estatísticas calculadas para testar o ajuste do modelo, tabela 4.2, e dos coeficientes da regressão, tabela 4.3, podemos perceber que apesar do modelo ser estatisticamente significante, $p < 0.0001$, temos que o menor $p-valor$ para os coeficientes é de 0.0647, que não é significante ao nível de 5%. Este resultado pode ser uma consequência

da presença de variáveis colineares. O modelo como um todo se ajusta bem aos dados, mas como algumas variáveis explicativas estão medindo fenômenos físicos similares, fica difícil determinar quais variáveis são importantes para o modelo.

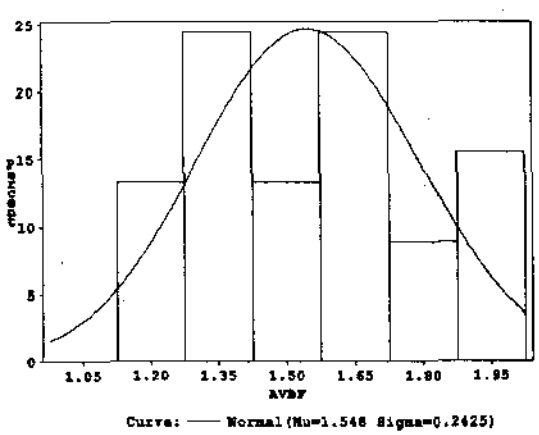
Antes de abandonarmos completamente o diagnóstico proposto por BKW, resolvemos infringir um pouco as recomendações, e mesmo não tendo um índice de condição maior que 30, optamos por avaliar a PDV. Percebemos, então, que para o índice de condição igual a 10.8, a proporção da variância é grande para as variáveis DEP e WTWAT, 0.7708 e 0.8029 respectivamente. O mesmo pode ser visto para o índice de condição igual a 14.1, onde a proporção de decomposição da variância é grande para as variáveis LWT e CWT, 0.9136 e 0.9629 respectivamente.

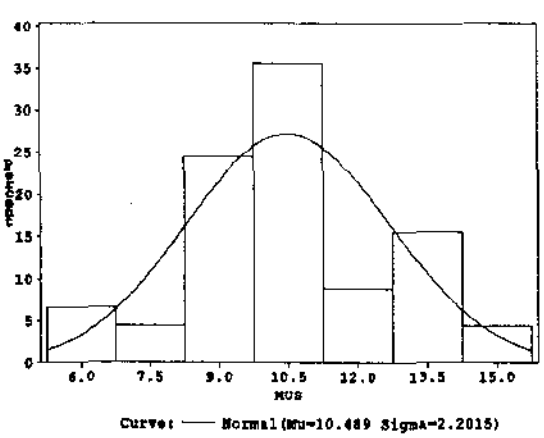
Diante desse fatos achamos que seria interessante refazer a análise, utilizando o diagnóstico $\bar{q}$ proposto.

Fizemos a princípio uma análise exploratória das variáveis que incluiu uma simples análise descritiva, com o teste para normalidade e os histogramas de cada uma das variáveis.

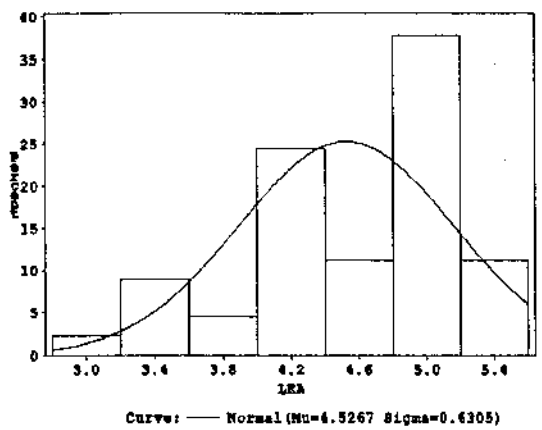
Pelo histogramas suspeitamos que algumas variáveis não tem distribuição Normal.

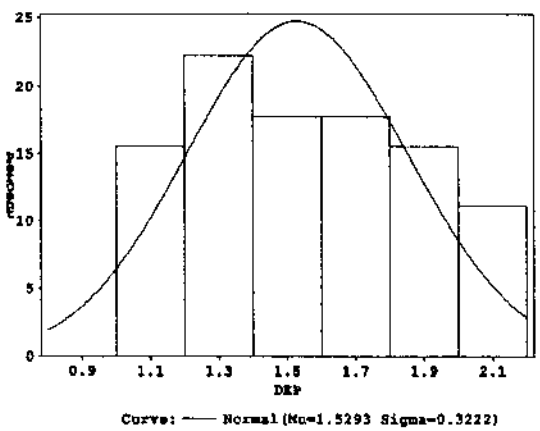
Histogramas para as variáveis AVBF, MUS.

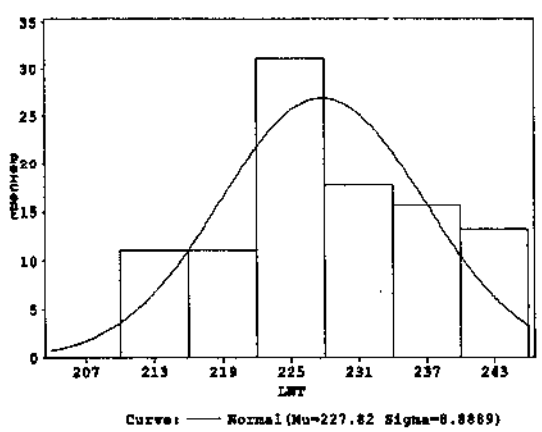


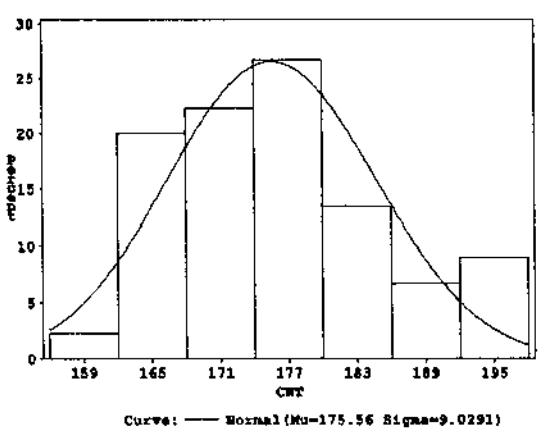


Histogramas para as variáveis LEA, DEP, LWT e CWT.

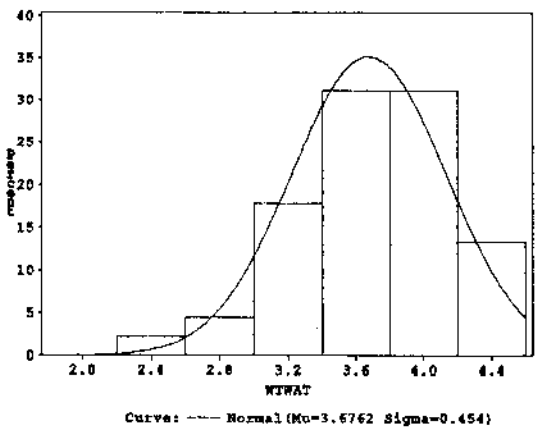


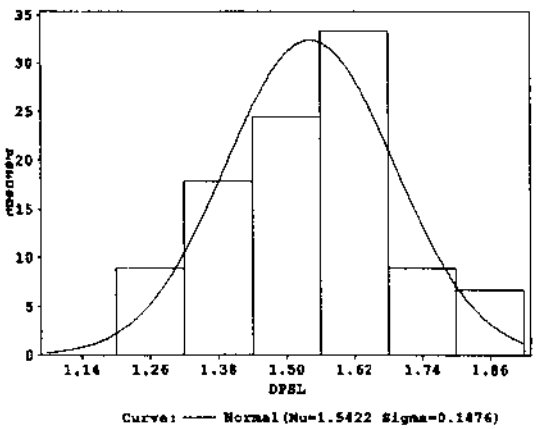


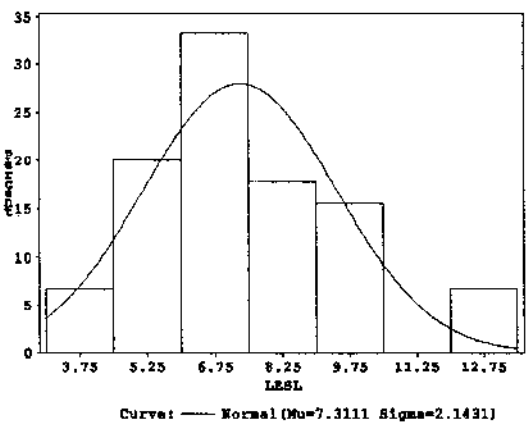


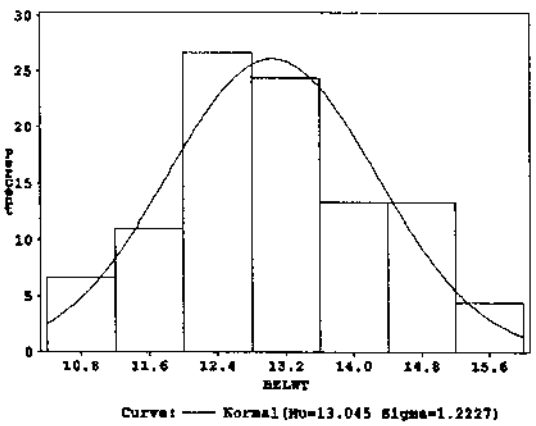


Histogramas para as variáveis WTWAT, DPSL, LESL e BELWT.









A suspeita inicial de não-normalidade para algumas variáveis é confirmada, como pode ser observado, na tabela 4.4, através do teste de Shapiro e Wilk (Est-W). As variáveis AVBF e WTWAT rejeitam a hipótese de normalidade ao nível de 5% e ainda as variáveis LEA e MUS apresentam $p - valores$ muito próximos de 5%.

Tabela 4.4

Variáveis	Média	D.P.	Est-W	p-valor
AVBF	1.5480	0.2425	0.9324	0.0150
MUS	10.489	2.2015	0.9458	0.0537
LEA	4.5267	0.6305	0.9451	0.0502
DEP	1.5293	0.3222	0.947	0.0637
LWT	227.82	8.8889	0.9640	0.2673
CWT	175.56	9.0291	0.9528	0.1017
WTWAT	3.6762	0.4540	0.9276	0.0094
DPSL	1.5422	0.1476	0.9718	0.4754
LESL	7.3111	2.1431	0.9528	0.1018
BELWT	13.045	1.2227	0.9652	0.2946

Decidimos por padronizar todas as variáveis e deixar de lado o fato de algumas não serem normais. Prentendemos com isso ver como a razão $\bar{q}$ se comporta com relação a não normalidado. Os resultados obtidos estão apresentados na tabela 4.5.

Tabela 4.5

Variáveis	**Razão** $\bar{q}(A_i^*/A_0^*)$
AVBF	2.3495
MUS	1.8689
LEA	2.4495
DEP	4.2023
LWT	5.1074
CWT	5.4083
WTWAT	3.3311
DPSL	1.4583
LESL	1.6585
BELWT	1.7586

Vemos através da razão $\bar{q}$ que as variáveis AVBF, LEA, DEP, WTWAT, LWT e CWT se destacam das demais e levantam a suspeita de serem variáveis colineares; poderiamos sugerir que: AVBF e LEA são variáveis colineares, $\bar{q}(A_1^*/A_0^*) = 2.3495$ e $\bar{q}(A_3^*/A_0^*) = 2.4495$, DEP e WTWAT também são variáveis colineares, $\bar{q}(A_4^*/A_0^*) = 4.2023$ e $\bar{q}(A_7^*/A_0^*) = 3.3311$ e ainda LWT e CWT são colineares, $\bar{q}(A_5^*/A_0^*) = 5.1074$ e $\bar{q}(A_6^*/A_0^*) = 5.4083$, concordando com a análise via proporção de decomposição da variância.

Os valores de $\bar{q}$ mostram que a colinearidade não é muito forte. Se considerarmos que no exemplo são mantidas as condições impostas nos modelos simulados no Capítulo 3, poderiamos dizer que a variável $X_5 = X_6 + 0.2\epsilon_1$, que $X_4 = X_7 + 0.25\epsilon_1$ e $X_3 = X_1 + 0.5\epsilon_1$ onde $\epsilon_1 \sim N(0,1)$, ou ainda suspeitarmos de uma colinearidade que envolve as variáveis X_5 e X_6 e outra envolvendo X_1, X_3, X_4 e X_7. Este é um dos casos que não foi estudado no nosso trabalho.

É importante salientar, que as análises feitas foram baseadas nos modelos estudados no Capítulo 3 e que quando novos estudos forem realizados, extendendo o que foi feito

neste trabalho, conclusões diferentes podem surgir. É este o motivo que nos leva a checar os resultados obtidos através do diagnóstico da razão $\bar{q}$, com os obtidos através da proporção de decomposição da variância, pois neste trabalho não esgotamos todas as diferentes situações que podem surgir, ao analisarmos um conjunto de dados.

4.2 Aplicação 2: Biomass

Este exemplo foi retirado de Rawlings (1988)[20] e tem como objetivo identificar as propriedades físicas e químicas dos substratos que influenciam na determinação da grande variedade de produção de biomassa *Spartina (*BIOMASS*)* em *Cape Fear Estuary.* Foram medidas catorze propriedades físico-químicas dos substratos do solo em uma amostra de quarenta e cinco locais. As variáveis são listadas à seguir:

H_2S: Sulfato Livre.

SAL: Salinidade.

EH7: Potencial de Redox no PH-7.

PH: PH na água.

BUF: Acidez do PH na solução tampão.

P: Concentração de Fósforo.

K: Concentração de Potássio.

Ca: Concentração de Cálcio.

Mg: Concentração de Magnésio.

Na: Concentração de Sódio.

Mn: Concentração de Manganês.

Zn: Concentração de Zinco.

Cu: Concentração de Cobre.

NH_4: Concentração de Amônia.

Utilizando os diagnósticos propostos por BKW, para o número e os índices de condição e a proporção de decomposição da variância obtemos os seguintes resultados:

Tabela 4.6

Autovalores	Índice de Condição	Prop. Var. H_2S	Prop. Var. SAL	Prop. Var. EH7	Prop. Var. PH
4.94348	1.00000	0.0018	0.0007	0.0013	0.0006
3.70334	1.15537	0.0001	0.0000	0.0102	0.0000
1.60203	1.75663	0.0056	0.0764	0.0499	0.0009
1.30627	1.94536	0.1329	0.0084	0.0318	0.0000
0.68940	2.67781	0.0001	0.1241	0.0210	0.0002
0.46837	3.24879	0.1094	0.0002	0.3565	0.0009
0.40512	3.49323	0.0343	0.0955	0.2312	0.0001
0.38519	3.58246	0.0450	0.0175	0.0091	0.0001
0.16557	5.46418	0.0646	0.0477	0.2038	0.0002
0.15114	5.71912	0.1685	0.1312	0.0647	0.0019
0.08861	7.46939	0.1915	0.0247	0.0007	0.0009
0.04960	9.98360	0.0003	0.0893	0.0000	0.0375
0.03214	12.4015	0.0001	0.1476	0.0106	0.0008
0.00975	22.5176	0.2457	0.2367	0.0093	0.9557

Obs: Continua na página seguinte.

Prop. Var. BUF	Prop. Var. P	Prop. Var. K	Prop. Var. Ca	Prop. Var. Mg	Prop. Var. Na
0.0010	0.0076	0.0000	0.0017	0.0001	0.0000
0.0000	0.0025	0.0094	0.0004	0.0030	0.0066
0.0009	0.0106	0.0001	0.0016	0.0001	0.0002
0.0005	0.0157	0.0002	0.0003	0.0001	0.0002
0.0010	0.4042	0.0013	0.0039	0.0010	0.0104
0.0000	0.0009	0.0000	0.0255	0.0000	0.0006
0.0000	0.3836	0.0068	0.0042	0.0011	0.0046
0.0028	0.0160	0.0042	0.0009	0.0035	0.0751
0.0013	0.0170	0.2709	0.0133	0.0002	0.0065
0.0024	0.0000	0.3190	0.0013	0.0012	0.1349
0.0361	0.0080	0.0335	0.1682	0.0143	0.0593
0.0979	0.0185	0.0160	0.1547	0.2550	0.3227
0.1598	0.0986	0.2705	0.0345	0.5454	0.1547
0.6962	0.0168	0.0683	0.5897	0.1750	0.2241

Obs: Continua na página seguinte.

Prop. Var. Mn	Prop. Var. Zn	Prop. Var. Cu	Prop. Var. NH_4
0.0024	0.0001	0.0037	0.0102
0.0016	0.0038	0.0006	0.0015
0.0001	0.0024	0.0096	0.2232
0.0316	0.0012	0.0014	0.2387
0.0001	0.0012	0.0000	0.1429
0.0495	0.0048	0.0067	0.0085
0.0581	0.0077	0.0512	0.0848
0.0407	0.0157	0.1187	0.1071
0.0027	0.0001	0.8029	0.0351
0.1349	0.9629	0.0051	0.1480
0.1595	0.3346	0.0976	0.0180
0.0087	0.0312	0.3633	0.1682
0.1790	0.5247	0.0172	0.2913
0.3309	0.0143	0.0344	0.1800

Analisando os resultados obtidos, vemos que o maior índice de condição é igual a 22.5 o que pela sugestão de BKW não representa uma colinearidade significativa, isto é, está associada a uma correlação entre as variávies inferior a 90%. Entretanto, se observarmos os resultados referentes ao ajuste do modelo linear, tabelas 4.7 e 4.8, percebemos que os resultados encontrados através dos Mínimos Quadrados foram afetadas pela colinearidade.

Tabela 4.7

		Análise de	Variância		
Fonte de Variação	**Graus de Liberdade**	**Soma de Quadrados**	**Média de Quadrados**	**F** -	$p-valor$ -
Regressão	14	15305139.1	1093224.2	8.436	0.0001
Erro	29	3757931.5	129583.8	-	-
Total	43	19063070.5	-	-	-

Tabela 4.8

Variáveis	**G.L.**	**Parâmetros Estimados**	**Erro Padrão**	**t para H_0: Parâmetro=0**	$Prob > \|t\|$
Intercepto	1	2603.31	3477.75	0.749	0.4602
H_2S	1	0.8531	3.0807	0.277	0.7838
SAL	1	-20.161	26.403	-0.764	0.4513
EH7	1	1.7389	2.0326	0.856	0.3993
PH	1	285.89	338.80	0.844	0.4057
BUF	1	0.7645	126.77	0.006	0.9952
P	1	-1.9841	2.7011	-0.735	0.4685
K	1	-1.0553	0.5048	-2.091	0.0454
Ca	1	-0.1303	0.1278	-1.020	0.3162
Mg	1	-0.2438	0.2801	-0.870	0.3912
Na	1	-0.0005	0.0257	-0.021	0.9835
Mn	1	-1.2726	5.4987	-0.231	0.8186
Zn	1	-18.304	22.586	-0.810	0.4243
Cu	1	329.66	113.23	2.911	0.0068
NH_4	1	-2.5976	3.3346	-0.779	0.4423

A regressão por Mínimos Quadrados Ordinários de Biomass nas catorze variáveis explicativas (listadas acima) levou a um $R^2 = 0.807$ com um modelo estatisticamente significante, $p < 0.0001$, apesar de somente as variáveis K e Cu serem significantes ao nível de 5%. Se avaliarmos os erros padrão associados às variáveis, percebemos que em alguns casos eles estão bastante elevados, por exemplo para as variáveis SAL, PH, BUF entre outras. Estes resultados podem ser decorrentes da presença de variáveis colineares no modelo, pois apesar do modelo como um todo se ajustar bem aos dados, as estimativas individuais podem ficar comprometidas. Além do que nada garante que este é o melhor ajuste, já que a possível presença da colinearidade prejudica a determinação das variáveis realmente importantes para o modelo.

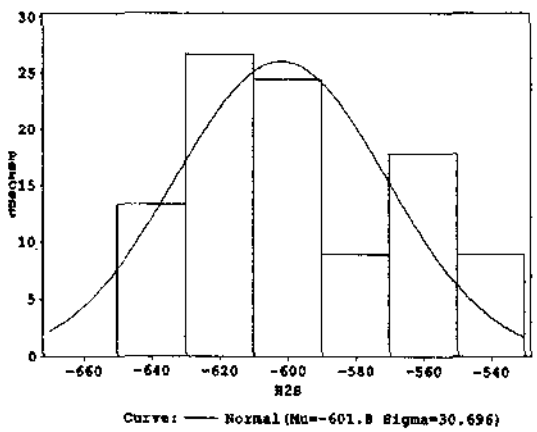
Diante desses fatos, achamos que seria interessante avaliar a proporção de decomposição da variância, mesmo o número de condição sendo menor que 30. Para o índice de condição igual a 12.4 as variáveis Mg e Zn apresentam a PDV associada de 0.5454 e 0.5247 respectivamente, e para o índice de condição igual a 22.5 as variáveis PH, BUF e Ca apresentam a PDV associada de 0.9557, 0.6962 e 0.5897 respectivamente, mostrando assim que existem dois conjuntos de variáveis colineares.

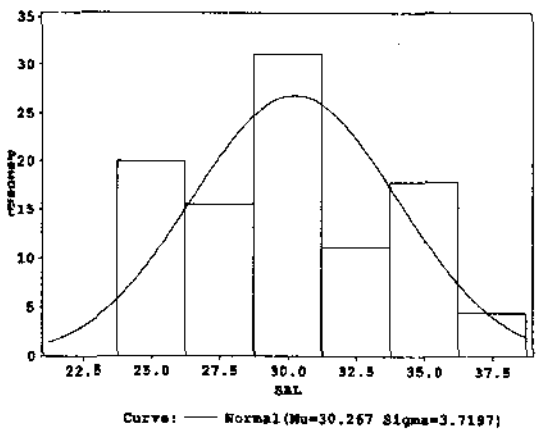
Decidimos então refazer as análises, utilizando o diagnóstico proposto $\bar{q}$.

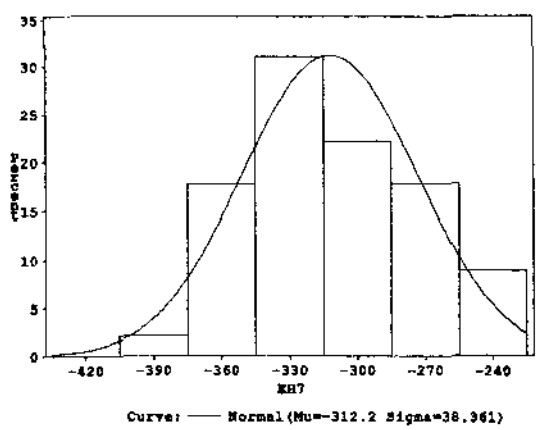
Fizemos, a princípio, uma análise exploratória de dados que inclui uma análise descritiva simples, com teste para normalidade e os histogramas para cada uma das variáveis.

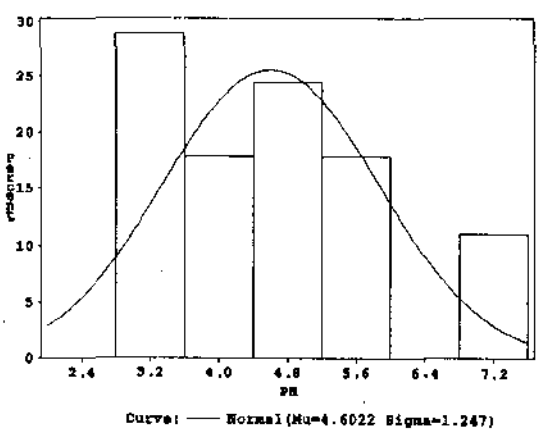
Pelos histogramas suspeitamos que algumas variáveis não tem distribuição Normal.

Histogramas para as variáveis H_2S, SAL, EH7, PH.

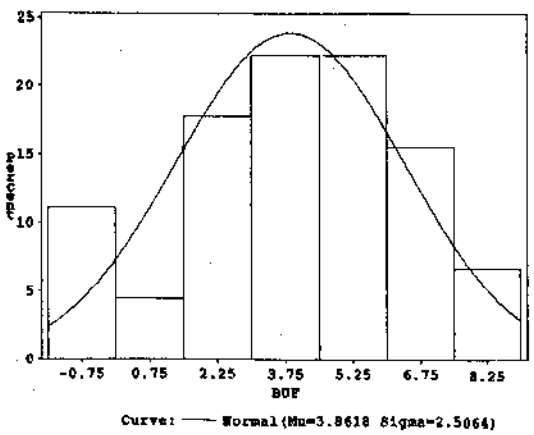


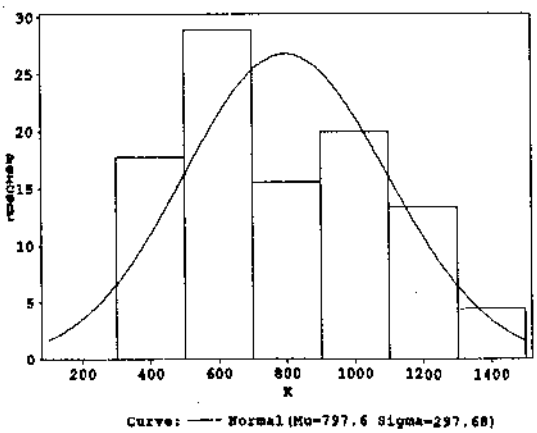


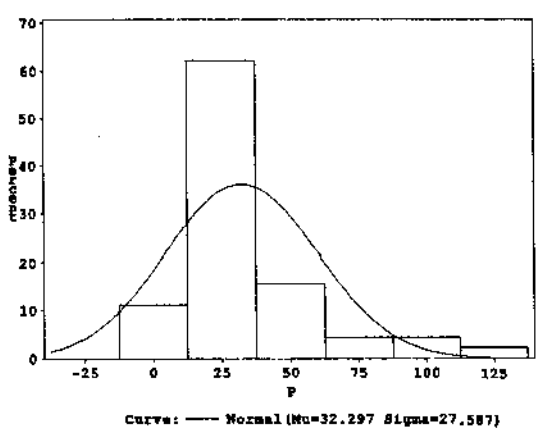


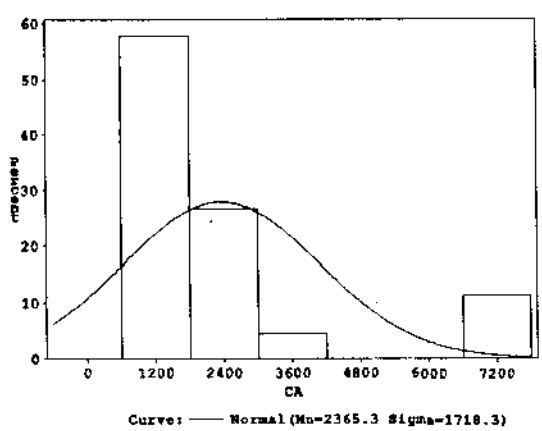


Histogramas para as variáveis BUF, P, K, Ca.

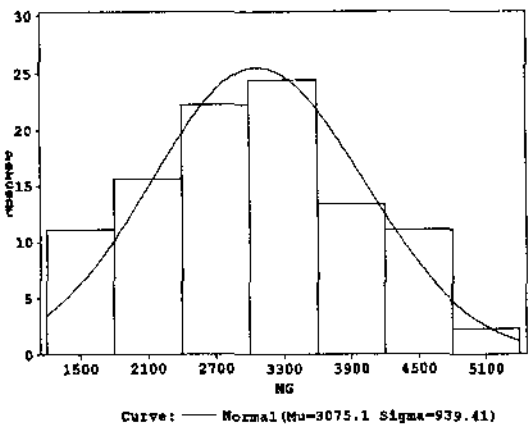


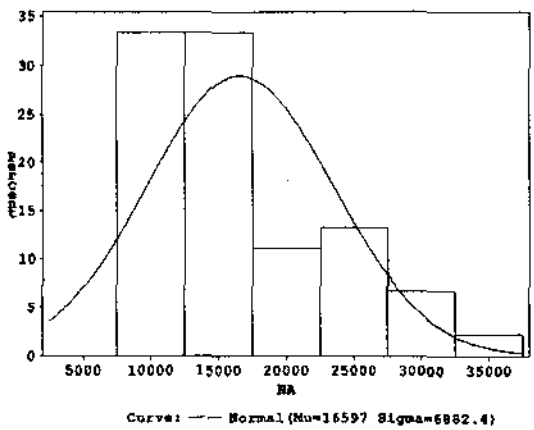


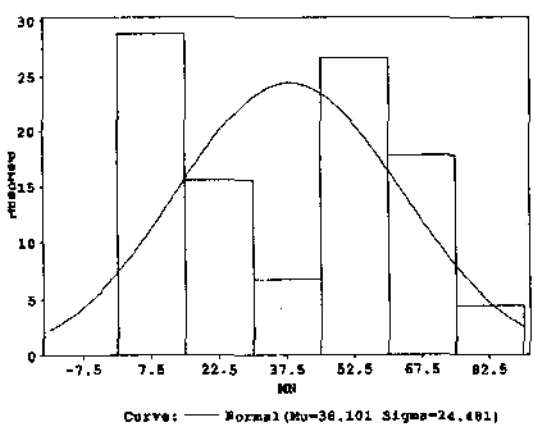


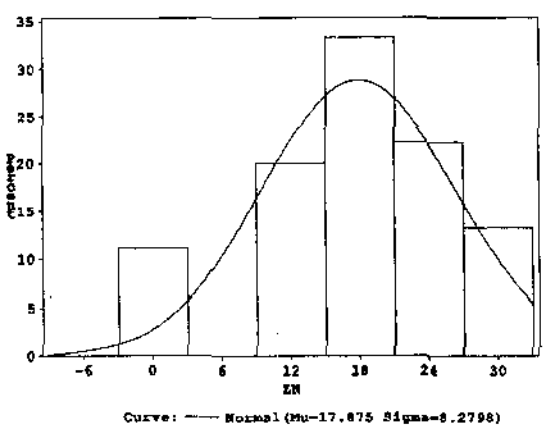


Histogramas para as variáveis Mg, Na, Mn, Zn.

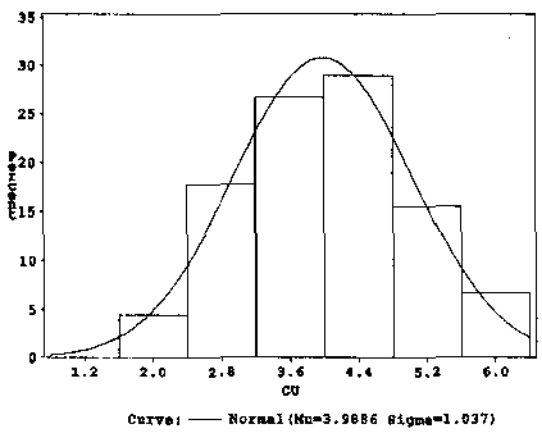


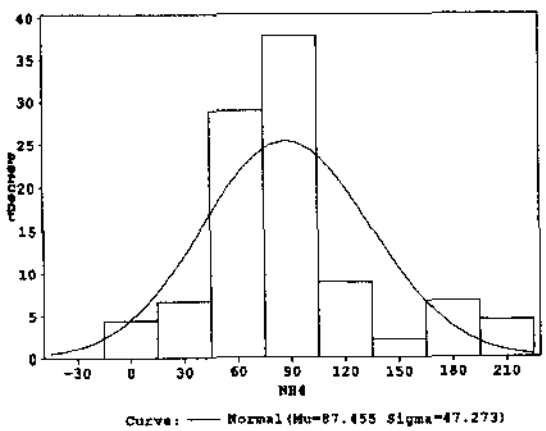






Histogramas para as variáveis Cu e NH_4.





A suspeita inicial de que algumas variáveis não são normais é confirmada. Como pode ser observado na tabela 4.9 através do teste de Shapiro e Wilk (Est-W), com exceção das variáveis EH7, BUF, Mg e Cu as demais rejeitam a hipótese de normalidade ao nível de 5%.

Tabela 4.9

Variáveis	Média	D.P.	Est-W	p-valor
H_2S	-601.78	660.08	0.8851	0.0002
SAL	30.267	3.7197	0.9411	0.0345
EH7	-312.18	38.361	0.9727	0.5063
PH	4.6022	1.2470	0.8701	0.0001
BUF	3.8618	2.5064	0.9567	0.1448
P	32.297	27.587	0.8047	0.0001
K	797.60	297.68	0.9132	0.0023
Ca	2365.3	1718.3	0.6210	0.0001
Mg	3075.1	939.41	0.9666	0.3264
Na	16596.7	6882.4	0.9096	0.0017
Mn	38.101	24.481	0.8843	0.0002
Zn	17.875	8.2798	0.9191	0.0041
Cu	3.9886	1.0370	0.9840	0.8772
NH_4	87.455	47.273	0.9107	0.0018

Para a aplicação do diagnóstico da razão $\bar{q}$ padronizamos as variáveis e desconsideramos o fato de algumas não serem normais. Os resultados obtidos estão na tabela 4.10.

Tabela 4.10

Variáveis	Razão $\bar{q}(A_i^*/A_0^*)$
H_2S	1.4211
SAL	1.7811
EH7	1.1818
PH	7.0850
BUF	4.7020
P	1.7168
K	4.1160
Ca	3.7272
Mg	5.0454
Na	2.6993
Mn	2.1117
Zn	3.1178
Cu	1.9465
NH_4	2.3776

Através da razão $\bar{q}$ vemos que as variáveis PH, BUF, K, Ca, Mg, Zn se destacam das demais por apresentarem valores mais altos. Entretanto não conseguimos, apenas com a razão $\bar{q}$, destacar os sub-grupos de variáveis que apresentam colinearidade. Neste caso a associação com a proporção de decomposição da variância é necessária, para descobrirmos quais os sub-grupos de variáveis estão envolvidas em colinearidade. Como no exemplo anterior, estamos considerando que os nossos dados se comportam como nos modelos simulados, isto pode estar prejudicando a eficiência do diagnóstico. Entretanto, este problema só poderá ser solucionado quando novos modelos forem simulados e mais situações exploradas.

Por exemplo, podemos suspeitar que o fato de muitas variáveis não terem distribuição Normal, ou pelo menos distribuição simétrica, está prejudicando a eficiência do di-

agnóstico, pois neste trabalho utilizamos o método percentil para o cálculo dos intervalos de confiança que é recomendado para distribuições simétricas, Efron (1993) [6]. Em estudos futuros é necessário a utilização de outros métodos para o cálculo dos intervalos de confiança quando a distribuição não é simétrica, como por exemplo o método do percentil ajustado.

Apêndice A

Programas

Os programas foram feitos utilizando o procedimento SAS/IML. Todos os arquivos de entrada são arquivos SAS com n linhas e $p+1$ colunas onde a primeira coluna é a variável resposta, y, e as restantes são as variáveis, X_i.

A.1 Coeficiente de Variação

```
proc iml ;
/* Programa para calcular as Medias dos Coeficientes De Variacao */
/* rotina para gerar as linhas a serem reamostradas */
start BOOT(Xstar,Ystar,X,Y,n);
do i=1 to n;
t=int((n)*UNIFORM(123457)+1);
if i>1 then Xstar=Xstar//X[t,];
else Xstar=X[t,];
if i>1 then Ystar=Ystar//Y[t];
else Ystar=Y[t];
end;
finish;
/* rotina para a estimacao dos Bs */
/* calculo dos coeficientes de variacao */
```

```
start ESTIMA(Bhatstar,CVstar,Xstar,Ystar,R,n);
Bhatstar=ginv(Xstar'*Xstar)*Xstar'*Ystar;
CVstar=sqrt(vecdiag((ssq(R)/(n-1))*ginv(Xstar'*Xstar))) /Bhatstar;
finish;
use ana.X12v1;
read all var{Y,X0,X1,X2,X3,X4,X5};
X=(X0||X1||X2||X3||X4||X5);
n=nrow(X);
p=ncol(X);
nboot=250;
/* replicacoes bootstrap */
Beta={4,2,5,-2,1,3};
do replic=1 to nboot;
run BOOT(Xstar,Ystar,X,Y,n);
run ESTIMA(Bhatstar,CVstar,Xstar,Ystar,R,n);
if replic>1 then Rstar=Rstar//Bhatstar';
else Rstar=Bhatstar';
if replic>1 then CV=CV//CVstar';
else CV=CVstar';
end;
/* Calculo das medias dos coeficientes de variacao */
CV1M=sum(CV[,1])/nboot;
CV2M=sum(CV[,2])/nboot;
CV3M=sum(CV[,3])/nboot;
CV4M=sum(CV[,4])/nboot;
CV5M=sum(CV[,5])/nboot;
CV6M=sum(CV[,6])/nboot;
print CV1M CV2M CV3M CV4M CV5M CV6M;
```

A.2 Índice e Número de Condição

```
proc iml ;
start BOOT(Xstarp,Ystar,Xp,Y,n);
do i=1 to n;
t=int((n)*UNIFORM(123457)+1);
if i>1 then Xstarp=Xstarp//Xp[t,];
else Xstarp=Xp[t,];
if i>1 then Ystar=Ystar//Y[t];
else Ystar=Y[t];
end;
finish;
start ESTIMA(icon,Xstarp);
autoxp=eigval(Xstarp'*Xstarp);
maxi=max(autoxp);
icon=(maxi/autoxp)##.5;
finish;
use ana.X12v2;
read all var{Y,X0,X1,X2,X3,X4,X5};
/* Padronizacao das Variaveis /*
X0p=X0/(SQRT(SSQ(X0)));
X1p=X1/(SQRT(SSQ(X1)));
X2p=X2/(SQRT(SSQ(X2)));
X3p=X3/(SQRT(SSQ(X3)));
```

```
X4p=X4/(SQRT(SSQ(X4)));
X5p=X5/(SQRT(SSQ(X5)));
Xp=(X0p||X1p||X2p||X3p||X4p||X5p);
n=nrow(Xp);
p=ncol(Xp);
nboot=250;
/* calculo dos indices de condicao */
somaicon=j(p,1,0);
do replic=1 to nboot;
run BOOT(Xstarp,Ystar,Xp,Y,n);
run ESTIMA(icon,Xstarp);
somaicon=somaicon+icon;
end;
medicon=somaicon/nboot;
print nboot medicon;
```

A.3 Proporção de Decomposição da Variância

```
proc iml ;
/* Programa para o Calculo da Proporcao de Decomposicao da Variancia */
/* Rotina para troca das linhas da matriz Xcp e das mesmas linhas do vetor Y*/
start BOOT(Xstarp,Ystar,Xcp,Y,n);
do i=1 to n;
t=int((n)*UNIFORM(123457)+1);
if i>1 then Xstarp=Xstarp//Xcp[t,];
else Xstarp=Xcp[t,];
if i>1 then Ystar=Ystar//Y[t];
else Ystar=Y[t];
end;
finish;
/* Rotina para calculo da PDV, para cada matriz Xstarp */
start ESTIMA(pi,Xstarp);
autoxp=eigval(Xstarp'*Xstarp);
autvexp=eigvec(Xstarp'*Xstarp);
p=nrow(autoxp);
fikj=j(p,p,0);
pi=j(p,p,0);
do i=1 to p;
fikj[i,]=autvexp[i,]##2/(autoxp');
pi[i,]=fikj[i,]/sum(fikj[i,]);
end;
```

```
finish;
use ana.X12v1;
read all var{Y,X0,X1,X2,X3,X4,X5};
/* Centralizacao da matriz X */
X0c=X0/(sum(X0)/50);
X1c=X1/(sum(X1)/50);
X2c=X2/(sum(X2)/50);
X3c=X3/(sum(X3)/50);
X4c=X4/(sum(X4)/50);
X5c=X5/(sum(X5)/50);
/* Padronizacao da matriz Xc */
X0cp=X0c/(SQRT(SSQ(X0c)));
X1cp=X1c/(SQRT(SSQ(X1c)));
X2cp=X2c/(SQRT(SSQ(X2c)));
X3cp=X3c/(SQRT(SSQ(X3c)));
X4cp=X4c/(SQRT(SSQ(X4c)));
X5cp=X5c/(SQRT(SSQ(X5c)));
Xcp=(X0cp||X1cp||X2cp||X3cp||X4cp||X5cp);
n=nrow(Xcp);
p=ncol(Xcp);
nboot=100;
somapi=j(p,p,0);
/* Replicacoes bootsratp */
do replic=1 to nboot;
run BOOT(Xstarp,Ystar,Xcp,Y,n);
run ESTIMA(pi,Xstarp);
somapi=somapi+pi';
end;
```

```
/* Calculo da media da PDV */
pdvstar=somapi/nboot;
print pdvstar;
```

A.4 Intervalos de Confiança

Este programa é referente a replicação bootstrap de tamanho 250, para as demais replicações os programas são análogos, sendo necessário fazer somente a adaptação para o número da replicação na linha "nboot=250" e os percentis nas linhas "l=7 e u=243 ".

```
proc iml ;
/* Programa para o calculo dos Intervalos de Confianca */
/* Amplitudes, Medianas e Media dos coeficientes bootstrap estimados */
/* Nboot=250 */
/* Rotina para gerar as linhas a serem trocadas */
start BOOT(Xstar,Ystar,X,Y,n);
do i=1 to n;
t=int((n)*UNIFORM(123457)+1);
if i>1 then Xstar=Xstar//X[t,];
else Xstar=X[t,];
if i>1 then Ystar=Ystar//Y[t];
else Ystar=Y[t];
end;
finish;
/* Rotina para estimar os coeficientes bootstrap */
start ESTIMA(Bhatstar,Xstar,Ystar);
Bhatstar=ginv(Xstar`*Xstar)*Xstar`*Ystar;
finish;
use ana.X12v4;
read all var{Y,X0,X1,X2,X3,X4,X5};
```

```
X=(X0||X1||X2||X3||X4||X5);
n=nrow(X);
p=ncol(X);
nboot=250;
Beta={4,2,5,-2,1,3};
/* Replicacoes boostrap */
do replic=1 to nboot;
run BOOT(Xstar,Ystar,X,Y,n);
run ESTIMA(Bhatstar,Xstar,Ystar);
if replic>1 then Rstar=Rstar//Bhatstar';
else Rstar=Bhatstar';
end;
/* Medias dos coeficientes estimados bootstrap */
B0hm=sum (Rstar[,1]/nboot);
B1hm=sum (Rstar[,2]/nboot);
B2hm=sum (Rstar[,3]/nboot);
B3hm=sum (Rstar[,4]/nboot);
B4hm=sum (Rstar[,5]/nboot);
B5hm=sum (Rstar[,6]/nboot);
print B0hm B1hm B2hm B3hm B4hm B5hm;

/* Intervalos de Confiancas, medianas, amplitudes */

do i=1 to 6;
  aux=Rstar[,i];
  k=aux;
  aux[rank(aux),]=k;
  if i>1 then B=B || aux;
```

```
else B=aux;
l=7;
u=243;
lm1=125;
sm2=126;
m1=B[lm1,i];
m2=B[lm2,i];
meaux=(m1+m2)/2;
if i>1 then me=me| | meaux;
else me=meaux;
print me;
icl=B[l,i];
icu=B[u,i];
ampaux=icu-icl;
if i>1 them amp=amp | | ampaux;
else amp=ampaux;
print amp;
```

Referências Bibliográficas

[1] Belsley, D.A., Kuh E. and Welsch, R.E. (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley and Sons, New York.

[2] Belsley, D.A. (1992) *Conditioning Diagnostics: Collinearity and Weak Data in Regression*, John Wiley and Sons, New York.

[3] Chatterjee, S. and Price, B. (1977) *Regression Analysis by Example*, John Wiley and Sons, New York.

[4] Draper, N.R. and Smith, H. (1981) *Applied Regression Analysis*, John Wiley and Sons, New York.

[5] Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife, *Annals of Statistics*, **7**, 1-26.

[6] Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*, Chapman & Hall, New York.

[7] Farrar, D.E. and Glauber, R.R. (1967) Multicollinearity in Regression Analysis: The Problem Revised, *Review of Economics and Statistics*, **49**, 365-366.

[8] Forsythe, G.E. and Moler, C.B. (1967) *Computer Solution of Linear Algebraic Systems*, Prentice-Hall: Englewood Cliffs.

[9] Golub, G.H. (1969) *Matrix Decomposition and Statistical Calculations*, Academic Press, New York.

[10] Gunst, R.F. and Mason, R.L. (1977) Advantages of Examinig Multicollinearity in Regression Anlysis, *Biometrics*, **33**, 249-260.

[11] Gunst, R.F. and Mason, R.L. (1980) *Regression Analysis and Its Application: A Data Oriented Approach*, Marcel-Dekker, New York.

[12] Gunst, R.F. (1983) Regression Analysis with Multicollinear Predictor Variables: Definition, Detection and Efects, *Communications in Statistics*, **12,** 2217-2260.

[13] Johnston, J. (1962) *Econometric Methods*, McGraw Hill, New York.

[14] Haitovsky, J. (1969) Multicollinearity in Regression Analysis: Comment, *Review of Economics and Statistics*, **50**, 472-479.

[15] Kendall, M.G. (1957) *A Course in Multivariate Analysis.* Griffin, London.

[16] Mansfield E.R. and Helms, B.P. (1982) Detecting Multicollinearity, *The American Statistician,* **36,** 158-160.

[17] Marquardt, D.W. and Snee, R.D. (1975) Ridge Regression in practice, *The American Statistician*, **29**, 3-20.

[18] Mason, R.L. , Gunst R.F. and Webster J.T. (1975) Regression Analysis and Problems of Multicollinearity, *Communications in Statistics*, **4,** 277-292.

[19] Montgomery, D.C. and Peck, E.A. (1982) *Introduction to Linear Regression Analysis*, Academic, New York.

[20] Rawlings, J.O. (1988) *Applied Regression Analysis: A Research Tool*, Wadsworth & Brooks/ Cole Advanced Books & Software, Pacific Grove, California.

[21] SAS Institute Inc. (1986) *SAS System for Regression,* Cary, NC, USA.

[22] Silvey, S.D. (1969) Multicollinearity and Imprecise Estimation. *Journal of the Royal Statistical Society*, Serie B, **31**, 539-552.

[23] Srivastava, M. and Sen, A. (1990) *Regression Analysis*, Springer-Verlag.